Faculty of Humanities

University of Copenhagen

Master's Thesis

Karina O'Neill

(ckw375@alumni.ku.dk)

The Automatic Identification of Dialogue Acts in Conversations

Supervisor: Costanza Navarretta (costanza@hum.ku.dk)

2017

ABSTRACT

Computational linguists study discourse structure to build a greater formal understanding of conversations, or discourse, and to build dialogue systems. There have been various attempts at modelling and automatically detecting discourse structure but there is at least partial agreement on the first level of analysis focusing on the identification of dialogue acts (DAs). The Switchboard corpus is a collection of human-human telephone conversations between pairs of speakers on a range of topics. Jurafsky et al. (1997) created a domain-independent tag set using this Switchboard corpus and in doing so created the Switchboard Dialogue Act Corpus (SWDA). For this study, I have looked at all DAs and then selected 10 of the most frequent DA tags in the SWDA to look at in more depth. I have attempted to compare a number of methods and features for the automatic identification of dialogue acts, including bag-of-words, random forests, decision trees and other classifiers. Some feature extraction was performed, particularly focusing on the tree root node labels and bag-of-words created from the utterances. There were two aims of the study. Firstly, an aim was to investigate the automatic identification of dialogue acts via comparison of machine learning classifiers. These classifiers were to be fitted to various text features data. Secondly, I aimed to explore the effect of parse tree root nodes as a feature on classification accuracy when the various classifiers were fitted to this feature. The prediction was that the various methods produce accuracies in the range of the baseline (39%) and the "gold standard" (84%).

The prediction was that the various methods produce accuracies in the range of the baseline (39%) and the "gold standard" (84%). Further work can continue to work towards a higher level of accuracy, for example by including acoustic data and other features of the SWDA corpus.

CONTENTS

INTRODUCTION

To build a greater formal understanding of conversations, or discourse, and to build dialogue systems, computational linguists describe and analyse discourse structure. Across many applications, as discussed below, the goal is to have a computer capable of accurately recognising, processing and responding to the natural language of humans. Discourse refers to a unit or units of spoken or written language utilised in a social context. A unit of speech, commonly ending with a change in turn by the speakers and a pause or silence, is referred to as an *utterance* (Loos et al., 2003). Discourse structure is "how a text is structured overall (i.e. how its parts are assembled)" (AQA.org.uk, 2015). It can be described in terms of a number of different levels:

> *Deep discourse structure*: At both the level of plans and intentions of the participants and the focus of attention in the conversation, linguists refer to modelling *deep discourse structure*. This is where traditional linguistics have focused more on (Moore & Wiemer-Hastings, 2003). Task-oriented automatic speech recognition systems more often integrate deep discourse structure into their language models, as this means specific discourse goals can be specified and a plan for achieving these goals can be integrated (Wright, 2000).
>
> *Shallow discourse structure*: At the level of speech act type of each utterance, or the level of sociolinguistic facts about the conversation, such as a participant's expectation of how the fellow participant will respond to a conversational unit, linguists refer to modelling *shallow discourse structure* (Jurafsky, Shriberg, & Biasca, 1997a; Stolcke et al., 2000). Shallow discourse structure is where computational linguistics and artificial intelligence are now focusing more on because it can be detected automatically with the computational tools available today. Wright (2000) noted that shallow discourse modelling is more often used for less task-specific speech recognition applications. It also allows for a larger data set to be analysed. This is the modelling approach that this study will take.

Computational linguists have attempted to model and automatically detect discourse structure, as this can allow a large amount of conversation data to be quickly and

systematically understood and can be used in applications such as dialogue systems, machine translation and automatic speech recognition (Král & Cerisara, 2014), which take natural language as input. A dialogue system is a computational system that interacts with humans using human language, either as text or speech. A well-known example of these applications is Apple's dialogue system app *Siri* (Schonfeld, 2010), acquired in 2010, which allows Apple device users to verbally communicate with their device using natural spoken language. Other large companies later introduced their own version of the personal voice assistant, to compete with the success of *Siri*, including Google's *Google Now* and Microsoft's *Cortana*. There is not yet consensus on a standard for discourse structure modelling and automatic detection, however Bunt et al. (2012) have attempted this with an ISO-published paper on a semantically-based standard for dialogue annotation. There are also numerous works on modelling and automatically detecting discourse structure that share at least partial agreement on what the first level of analysis should be - identifying dialogue acts.

A *dialogue act (DA)* represents the communicative function, or illocutionary force (speaker's intention), of an utterance in a conversation (Eckert & Strube, 2000; Stolcke et al., 2000). This is similar to the speech act defined by Searle (1969), which he described as the basic or minimal unit of linguistic communication that is employed with a certain intention. He listed some of these speech acts to include making statements, giving commands, asking questions and making promises. Across different dialogue act tag sets, the name given to "dialogue acts" varies. For example, in the Map Task corpus (Anderson et al., 1991), these are called "move types" because the corpus is from a task in which participants discussed moving around a map. In general, dialogue acts are semantic and pragmatic units that can be recognized in dialogues. The Guidelines for Dialogue Act and Addressee Annotation Version 1.0 (AMI Consortium, 2005), a manual written for annotation of the AMI (Augmented Multi-party Interaction) corpus, described how to annotate discourse with dialogue acts. It described dialogue acts as being defined by speaker intention, i.e. what the speaker is trying to achieve with what they are saying. It noted that dialogue act annotation is about function (what the person means) and not form (how they say it), though form can often give hints about what the function is. For example, question dialogue acts often start with a "Wh" word and speech pitch often goes up at the end of a question. The process of deciding where one dialogue act ends and the next begins is termed *segmentation*. As Byron and Stent (1998)

note, deciding on utterance boundaries can be difficult in spoken language, as annotation depends on criteria other than punctuation. However, once conversations have been transcribed, this syntactic information can be obtained from the written transcripts and this can help with deciding on utterance boundaries. It should be noted here that dialogue act annotation does not always take into account acoustic information. Stolcke et al. (2000) noted that excluding acoustic information increased the amount of data that could be covered by annotators. Shriberg et al. (1998) tested the effect of excluding acoustic information, by labelling a subset of data with acoustic information, and found that labelling accuracy only slightly improved. For most dialogue acts, not more than 2% of labels were changed by including the acoustic information.

Aims: There were two aims of the study. Firstly, an aim was to investigate the automatic identification of dialogue acts via comparison of machine learning classifiers. These classifiers were to be fitted to various text features data. Secondly, I aimed to explore the effect of parse tree root nodes as a feature on classification accuracy when the various classifiers were fitted to this feature. The prediction was that the various methods produce accuracies in the range of the baseline (39%) and the "gold standard" (84%).

The Background Literature section introduces a number of dialogue act tag schemes that have been proposed in previous work, as well as introducing the dialogue act tag set used in this study. Literature is then described for studies that have used classifier algorithms for the automatic identification of dialogue acts from corpora. In the Data section, a number of different corpuses that have been annotated with dialogue acts are introduced, with a deeper description and explanation of the corpus chosen for this study. Next, the Methodology section describes how the corpus data was modified for the purpose of this study and the method used for the automatic identification of dialogue acts from this data is described, namely the classification process. The evaluation of performance of this method is also described. Next, the Results and Analysis section describe and analyse the experimental results. The Discussion section covers the strengths, limitations, issues and potential improvements of the study and discusses future work to follow from this study. Finally, the Conclusion summarises the study findings and provides concluding comments. All referenced tables are given in the *Appendix* section.

BACKGROUND LITERATURE

Dialogue Acts

A number of dialogue act tagging schemes have been proposed and studied, being constructed based on different corpora and across domains. Here, some of them are described. The dialogue act scheme then chosen for this study is the DAMSL dialogue tag set, which is described in more detail below, specifically describing some of the most frequently occurring dialogue act tags from this set.

Austin (1962) was the first to refer to "speech acts", which he also called "performative utterances". He described these utterances as not only being used to say something but also to carry out some action. He divided these acts into several categories, which Searle (1976) improved on with his five "speech act" categories - *representatives, directives, commissives, expressives, declarations*. Each category encompasses a number of speech acts, for example *commissives* are those that commit the speaker to some future action, such as *promising* and *planning*. Walker and Whittaker (1990) focussed on creating a domain-independent scheme and on the shift of control between speakers in a conversation. They specified just four "utterance types" - *assertions, commands, questions* and *prompts* - which is much less than those proposed by Searle. Traum and Hinkelman (1992) proposed "conversation acts", which they described as building on the traditional speech acts to also include turn-taking, grounding (the process of achieving mutual understanding between conversation participants), and higher-level argumentation (the process of systematic reasoning) acts. Carletta et al. (1997) built the "move types" coding scheme from the Map Task, as described in the *Data* section, which has 12 dialogue acts. This scheme was built from task-oriented dialogue and the 12 acts are very domain-specific dialogue, so there is not great generalisability to general dialogue. It does not include social functions acts, as many of the other tagging schemes do. Jekat et al. (1995) and Shriberg et al. (2004) constructed annotation schemes for the VERBMOBIL and ICSI Meeting corpora, respectively, as described in the *Data* section. As with Carletta et al. (1997)'s scheme, these two schemes are domain-specific and task-oriented. As with Searle's speech act categories, both schemes have categories that divide into subcategories. For example, Jekat's higher-level act *clarify* divides

into *clarify_state*, *clarify_ answer* and *clarify_state*. Both schemes divide into more categories than described by Searle.

There is not yet one standard, agreed upon dialogue act scheme, with one criticism being that this is difficult to achieve because dialogue act taxonomies are based on intuition rather than statistical analysis of utterance classification accuracy (Andernach, Poel, & Salomons, 1997). It is also a problem that dialogue act schemes have so often been built from task-oriented corpora.

Frequent Dialogue Act Tags

As mentioned, the DAMSL dialogue act tag scheme was chosen for this study. It is a generalisable annotation scheme and it allows an utterance to have more than one act tag, as it acknowledges that more than one action can be performed with one utterance. It also allows for subcategories, so can be compared with other schemes at both the higher and lower levels of categorization (Hacquebord, 2017). To help explain further what dialogue acts are, I will describe some of the more frequent dialogue tags in the Switchboard corpus, as listed in Table 1. Jurafsky, Shriberg, and Biasca (1997a) provide a full description of the full DAMSL dialogue tag set. The tag abbreviations given in brackets do not always correspond directly to the tag names, as the names have changed slightly from the original pre-clustered tags. For example, the *statement-opinion* tag's abbreviation originates from the tag name *viewpoint*.

> *Statement-non-opinion (sd) and Statement-opinion (sv):* a distinction is made between non-opinion and opinion statements, with an sd described as "descriptive/narrative/personal", while an sv is "other-directed opinion" statements. A statement (sd) more often is responded to with a continuer/backchannel, i.e. is not disputed, while an opinion (sv) is often responded to with a further opinion, such as a disagreement, i.e. is disputed.
>
> *Acknowledge (Backchannel) (b):* acknowledgement by a speaker that the other speaker said something, for example acknowledging that the other speaker answered their question.

Agree/Accept (aa): utterance that signals the speaker's agreement with or acceptance of the utterance(s) of the previous speaker. The speaker agrees with or accepts the proposal, plan, opinion or statement being communicated by the previous speaker.

Appreciation (ba): a verbal expression of appreciation for something, similar to an acknowledgement but with a slightly higher level of emotional involvement and support.

Yes-No-Question (qy): a question that is pragmatically (i.e. functions as a question), syntactically and prosodically a yes-no-question.

Non-verbal (x): an utterance with non-verbal material, such as laughter and coughing.

Yes answers (ny): a positive answer that is either "yes" or a variant of this, i.e. "yeah", "uh-huh", "yep" etc. This can include answers with a pause or discourse marker together with the "yes", for example "well, yes".

Conventional-closing (fc): an utterance that closes a conversation, for example "Bye, nice talking to you."

Wh-Question (qw): a question that begins with a "wh-word", for example "What", "Who", "When".

Dialogue acts effectively form a tag set to classify utterances based on "pragmatic, semantic, and syntactic criteria" and this tag set is then used to build a discourse language model (Stolcke et al., 2000). These tag sets can either be domain-specific, where they are built for use with a specific application, or they can be domain-independent, to be used for broader application. The Discourse Annotation and Markup System of Labeling (DAMSL) tag set was an attempt at building one such domain-independent dialogue act tag set (Core & Allen, 1997). Jurafsky et al. (1997a) soon after created an augmented tagset by using the large Switchboard (SWBD) corpus (Godfrey & Holliman, 1997) and refining their language

model. This is described in the *Switchboard Dialog Act Corpus* section. A corpus is a large and structured set of texts used for linguistic research and analysis.

Machine Learning Classifiers

Machine learning classifiers are a type of predictive modelling technique. They take input variables and, from this, predict a target variable. In the example of this study, they would predict the dialogue act based on features of corpus data. *Features* are informative derived values from data. Numerous classification algorithms have been used in dialogue act automatic identification studies, some of which are described and then implemented in this study. Classifiers are not often compared in dialogue act recognition research and that if classifiers are compared, this will only involve two-to-three classifiers. Artificial neural networks have only recently begun to be used for classification in dialogue act recognition. Therefore, this study has attempted to compare numerous classification algorithms, including an artificial neural network, the multilayer perceptron (MLP). All classifiers described here are supervised classifiers. This means that the classifier initially receives training input data paired with their target variables. From this, they learn and make a function that can then make predictions about target variables when receiving unlabelled input data.

Decision tree classifier

A decision tree is a flowchart that takes some data with features as input and can predict a target label, or "class", for the input data. Breiman et al. (1984) proposed two types of decision trees - classification and regression trees (CART). A classification tree predicts what class the input data belongs to, for example gender or high-risk insurance customer, whilst a regression tree predicts a continuous value - a real number, for example age in years or house price. In this study, a classification tree was tested. The tree observes the value of a feature at branches, or "decision nodes", and then selects the next branch based on a condition set for a feature's value. Here, another decision node is reached, where another decision is made based on another feature condition. This process continues until the "leaf node" is reached. Here, a predicted target label, or class, for the input data is provided. A classification decision tree is able to work with multiple classes, i.e. it is not limited to binary classification. The deeper the tree is, the more complex the decision rules are.

Random forest classifier

A random forest, like a decision tree, can do either classification or regression. In this study, the classification type of random forest was tested. The random forest trains numerous decision trees on sub-samples of the training data and then the predicted class is calculated as the mode of all classes predicted by the decision trees. Ho (1995) was the first to create a random forest algorithm and Breiman (2001) developed an extension that is included in the implementation used in this study. A random forest classifier has the advantage of controlling for overfitting and improving prediction accuracy (Scikit-learn, 2011).

Majority classifier ('most frequent')

The majority classifier, implemented in this study using Scikit-learn's Dummy Classifier, is a simple classifier used only to give a baseline result to help compare other classifiers. It is not a classifier to be used for real-world problems. From the training data, the classifier finds the most frequently occurring class and then simply predicts that all test data will belong to this class. Therefore, the accuracy of this classifier is reflective of what proportion of the training data belongs to the most frequently occurring class. If 60% of the data belongs to this class, then the classifier will have a 60% classification accuracy, for example. In this study, this accuracy will reflective the frequency of the most frequently occurring dialogue act tag in the training data.

Multi-Layer Perceptron (MLP) classifier

A multi-layer perceptron is a type of artificial neural network (ANN). An ANN is based on the structure and functioning of biological nervous systems, notably the brain, as they process information. An ANN consists of interconnected elements, or 'nodes', that are working together to solve a problem, mirroring the interconnected neurons and synapses of the brain that pass information along each other. As in the biological nervous systems, ANNs go through a learning process to become competent at operations such as pattern recognition and (data) classification. In the nervous system, adjustments are made to synapses during and as a result of the learning process. ANNs receive these adjustments too, to their interconnected nodes.

A multi-layer perceptron is a feedforward artificial neural network, that uses *backpropagation* during training. Backpropagation is the process of adjusting the weight of each node by calculating the amount of error at the output compared to the expected output, after each data sample is passed through the network, and then updating the weights back through the network. The multilayer perceptron has at least three layers - the input layer, at least one hidden layer, and then an output layer. The nodes of the hidden layers use a non-linear activation function that learns via backpropagation. A nonlinear activation function is the process that takes an input and gives an output, with the input values passing through a nonlinear function during the process. At each layer of the model, the values of the input are modified using this function, until the output layer, where the values are finally transformed into the output value. As with other classifiers discussed, the MLP can do both regression and classification. In this study, the MLP will be used for classification, so the output values will be the predicted dialogue act.

(Multinomial) Naive Bayes classifier

A naive Bayes classifier is based on Bayes' theorem with the assumptions that input data features are independent from each other when contributing to the predicted class, despite any correlations between the features. Naive Bayes classifiers have worked quite well for some complex real-world problems, however it has also been shown to be outperformed by other classifier methods, such as random forests (Pranckevicius & Virginijus, 2017). There are a few different naive Bayes classifiers; this study tested the multinomial naive Bayes classifier. This classifier takes input data features with multinomial distribution. It usually takes integer feature counts as input, so is commonly used for text classification and so was suitable for this study. Although this classifier has been widely used, as described later, a disadvantage of using it is that it has been found to be a poor estimator.

Support Vector Machine (SVM) classifier ('one-vs-rest' scheme)

A support vector machine constructs a hyperplane, which is an n-1-dimensional space that is a subspace of a n-dimensional space, using training data points. The hyperplane divides the space, with one category, or class, on one side of the divide and a second category on the other side of the divide. Each new data point is mapped into the space either side of the hyperplane and in doing so, updates the hyperplane to be optimal for the training data so far

inputted. The optimal hyperplane will have the largest distance from the nearest data point of either class. So far, binary classification is described. In this study, there are 42 classes, so a classifier needs to be capable of multi-class classification. For this, a one-vs-rest scheme was utilized, meaning a support vector machine classifier was run for each class, i.e. 42 SVMs, fitting the class against the rest of the classes for each classifier. Compared to other SVMs, this scheme is computationally efficient. The one-vs-rest scheme is the most often used strategy for multiclass classification (Scikit-learn, 2011).

Classifiers in Previous Studies

It is difficult to compare results from previous studies that have used these classifiers, as there are many variables that vary across studies and affect accuracy. This includes different amounts of data and domains from which the data is drawn, different languages, and different dialogue act tag schemes used. We can, however, look at studies with similar setups and also studies which compared classifiers within the study.

Classifiers that have been used in automatic dialogue act classification include decision trees (Moldovan, Rus, & Graesser, 2011; Samei et al., 2014; Dbabis et al., 2015), neural networks (O'Shea, Bandar, & Crockett, 2012; Khanpour, Guntakandla, & Nielsen, 2016; Tran, Zukerman, & Haffari, 2017), naive Bayes (Kim, Cavedon, & Baldwin, 2010; Moldovan, Rus, & Graesser, 2011; Samei et al., 2014; Dbabis et al., 2015) and support vector machines (Surendran & Levow, 2006; Tavafi et al., 2013; Dbabis et al., 2015). Also there appears to be very few, perhaps only one, implementation of a random forest (Tran, Zukerman, & Haffari, 2017).

Both Moldovan, Rus and Graesser (2011) and Samei et al. (2014) tested a naive Bayes and a decision tree classifier on an online chat corpus. Moldovan, Rus and Graesser (2011) used the NPS Internet Chatroom Conversations corpus, annotated with the 15 speech act categories constructed by Forsyth and Martell (2007). Samei et al. (2014) used a corpus of internet gaming chats annotated with the 7 speech act categories that were proposed by Moldovan et al. (2011). Both studies found the decision tree classifier to perform better than the naive Bayes classifier, with a 5% increase from Moldovan et al. (2011) (they reached a

decision tree accuracy of 78.35%) and a 1.43% increase from Samei et al. (they reached a decision tree accuracy of 56.19%).

Dbabis et al. (2015) tested a support vector machine, as well as naive Bayes and decision tree classifiers, using Arabic human-human transcribed TV debates with a tag set relevant to argumentative conversations. Although they found relatively low accuracy results compared to other corpora, it can still be noted that they found the lowest accuracy with the decision tree classifier (F-measure = 24.3%) and highest with the SVM (F-measure = 29.8%), so these results are not in line with the two previous studies described.

The type of SVM used can affect accuracy. When a dataset is more than a couple of 10,000 samples, the data fitting time becomes very long, so a linear SVM can be more suitable. Model complexity is reduced with this model, which suggests the accuracy may also be affected. Surendran and Levow (2006) tested a linear SVM on the Map Task corpus. They found a 58.1% accuracy, though they were selective with features used. They used a bag-of-n-grams model, using only unigrams, bigrams and trigrams that appeared at least twice in the data, plus unigrams for words that were the only word in a dialogue act. Due to the specific, selective model used, it is difficult to compare with SVM results from other studies.

Tavafi et al. (2013) tested a multiclass SVM model on conversations across a variety of conversational domains - emails, forums, meetings and telephone conversations. For the telephone conversations data, the Switchboard corpus was used (Jurafsky et al., 1997). They used a reduced tag set size, compared to that of Jurafsky et al. (1997), which was a set of 19 tags from Jeong (2009). For the Switchboard data, the multiclass SVM performed only 0.06% better than the baseline of 46.44%, when micro-averaging was used, and performed the same as baseline, 6.25%, when macro-averaging was used (macro-averaging averages the results of each class, disregarding potential importance of class size, whilst micro-averaging weights results by class size).

The Loqui dialogue corpus (Passonneau & Sachar, 2014, described below) was also tested with a regular SVM, though this involved some more complex features, including preceding dialogue act tag predictions and DFUs (links between utterance pairs, described more in the *Data* section). Compared to a baseline of 50.14%, which is similar to that in Tavafi et al. (2013)'s study, the SVM accuracy was 68.3%. This gives some support to using corpora of telephone conversation transcriptions for automatic dialogue act classification.

Tran, Zukerman and Haffari (2017) tested both a recurrent neural network (RNN) and a random forest classifier on the Map Task corpus and achieved accuracies of 61.6% and 52.5%, respectively. This was using Carletta et al. (1997)'s 12 dialogue act tags plus an "unclassifiable" tag. This tag can be equated with Jurafsky et al. (1997)'s tag "uninterpretable". They also tested the recurrent neural network on the Switchboard corpus and achieved a 74.5% accuracy. This was using Jurafsky et al. (1997)'s 42 dialogue act tags. So, at least with the task-oriented Map Task corpus, a neural network performed better than a random forest classifier, though a random forest classifier is still worth utilizing for dialogue act classification. The success of the recurrent neural network was attributed to the fact that the model used both utterance-level and dialogue-level (long-range) dependencies. Dependency is the idea that linguistic units are connected by directed links. These links are called dependencies. For this study I will not look at dependencies, as it is still possible to achieve good accuracies using features other than dependencies, but this can be a consideration for future work and improvement. O'Shea, Bandar and Crockett (2012) created their own corpus from online FAQ lists and blogs and tested numerous classifiers and found good accuracies for multilayer perceptron (69.14%) and naive Bayes classifiers (55.98%), which adds these two classifiers to the dialogue act classification literature. However, the study only had two dialogue tags - question and non-question.

There have been various accuracies achieved with the ICSI Meeting Corpus, depending on how many dialogue act tags were used. These include 89.3% achieved with 5 tags (Verbree, Rienks, & Heylen, 2006), 80.5% with 11 tags (Tavafi et al., 2013) and 66% achieved with 62 tags (Ji & Bilmes, 2005). This shows that focusing on less tags can increase accuracy.

Previous studies have found relatively high labelling accuracy with classification algorithms when using the Switchboard corpus, the corpus that was modified for use in this study, which will be discussed later. For example, Stolcke et al. (2000) found a 71% labelling accuracy with hand-transcribed (true) words when using a combination of word-based likelihoods, which are the probabilities of words given the utterance DA tags, and a trigram discourse grammar. This is relatively high when compared to a baseline (chance) accuracy of 35% and human labelling accuracy of 84%. This suggests the corpus DA tags are relatively reliable. The human labelling accuracy is also referred to as the "*gold standard*" and it is achieved by having multiple linguistic experts annotate the corpus independently and then an

inter-annotator agreement is calculated, which controls for annotator bias. This is the gold standard. Khanpour, Guntakandla and Nielsen (2016) trained a long short-term memory (LSTM) RNN for DA classification using the Switchboard corpus and Jurafsky et al. (1997)'s 42 dialogue act tags. A LSTM is a RNN modified to maintain long-distance dependencies as their default specificity. They achieved an accuracy of 80.1%. This shows the potential of machine learning classifiers with the Switchboard corpus and this tag set, as this accuracy is coming close to the gold standard of 84%.

DATA

There are a number of different corpora used for dialogue act identification research and there are different types of corpora among these. There are human-human or human-computer dialogues, dialogues produced in natural or unnatural settings, dialogues produced from fictional works, and different modalities - written, spoken or face-to-face. Spoken and face-to-face dialogues have the additional auditory and visual information to help identify the correct dialogue act. However for this study, I have chosen to focus solely on one modality - word transcripts of spoken conversations. Here, the dialogue audio is recorded and then later transcribed.

Researchers either choose to annotate a corpus with dialogue act tags themselves, to be used in their research, or they use a pre-annotated corpus. The advantages of using a pre-annotated corpus are that it saves time, as manual annotation is laborious, and you can then compare your results to those of other studies.

Below is a description of some previously-used corpora. The corpus chosen for this study is the Switchboard Dialog Act corpus. The corpus data was then modified to serve the aims of this study, as described in the Methodology section.

Map Task Corpus

The Map Task corpus (Anderson et al., 1991) consists of 128 human-human task-oriented transcribed spoken dialogues. Pairs of participants had to verbally help each other with a mapping task. One participant had a route on a map and had to explain the route to the other participant who had to reproduce it on their map. It was annotated by Carletta et al. (1997) with a coding scheme of 12 dialogue acts or "move types". There is the limitation here of dialogue being constrained to the mapping task, so the dialogue acts do not cover social functions such as empathy, for example, instead focusing more on acts such as instruction and explanation.

VERBMOBIL Corpus

The VERBMOBIL corpus (Burger et al., 2000) consists of 726 human-human task-oriented transcribed spoken dialogues and is in English, Japanese and German. Participants were paired and had the task of arranging appointments. Jekat et al. (1995) annotated the corpus with 18 higher-level dialogue act tags and these were further divided into subcategories. For example, the higher-level act *clarify* divides into *clarify_state*, *clarify_ answer* and *clarify_state*.

ICSI Meeting Corpus

The ICSI Meeting corpus (Janin et al., 2003) contains 75 human-human transcribed spoken dialogues of meetings between groups of up to 6 people who were given specific topics to speak on related to linguistics. Shriberg et al. (2004) annotated the corpus with the Meeting Recorder Dialog Act (MRDA) tag set. As similar to the VERBMOBIL annotation tags, this tag set contains 11 general tags, which further divide into 39 specific tags.

NPS Internet Chatroom Conversations

The NPS Internet Chatroom Conversations corpus (Forsyth & Martell, 2007) contains 15 human-human online chatroom dialogues. The difference from the other corpora discussed is that the dialogues are computer-mediated written dialogues, which have differences from traditional written language such as the use of emoticons and abbreviations. Wu et al. (2005) annotated the corpus with 15 dialogue acts, which include acts such as *emotion* to cover the social nature of the dialogues.

Loqui Dialogue Corpus

The Loqui Dialogue Corpus (Passonneau & Sachar, 2014) contains 82 human-human transcribed spoken dialogues from telephone conversations between librarians and patrons. It has been annotated by Passonneau and Sachar with 13 dialogue acts, for example *inform, affirmative, backchannel.* They also annotated the corpus with Dialogue Function Units (DFUs), which links utterances, for example into adjacency pairs of questions and answers.

Switchboard Dialog Act Corpus

The data used for this study is augmented from the Switchboard corpus mentioned in the *Dialogue Acts* section above. The Switchboard corpus is a collection of 2,438 human-human telephone conversations in between 543 speakers in 1990 in the USA (Godfrey & Holliman, 1997). The demographic of the speakers was spread across ages 20-69, all education levels and covered every major dialect of American English. This broad demographic, compared to the subjects used in other corpora, adds some level of generalisability to findings from studies using this corpus. Speakers were recruited both internally, from the technology company (Texas Instruments) that recorded the telephone conversations, and externally, from members of the public including some involved in speech research. They were given an option to receive a financial reward ($5 for each completed call) for their participation. There were 70 possible topics that pairs of speakers could talk about. A computer-driven robot operator system selected the caller, callee and topic of conversation and then the conversations were recorded. Apart from the topic prompt, the conversations between speakers were spontaneous. This deliberate lack of interruption from the experimenters, and use of a standardised computer system across the data collection period, reduced experimenter bias. The conversations were rated by the transcribers as being highly natural, according to the corpus' user manual.

In the process of creating the augmented tagset mentioned above, Jurafsky, Shriberg and Biasca (1997b) created the Switchboard Dialogue Act Corpus (SWDA), which I have chosen to analyse for the purpose of this study. The SWDA is a subset of the SWBD corpus, containing 1,155 conversations that are annotated with utterance-level dialogue act tags. These tags are taken from the SWBD-DAMSL annotation scheme of Jurafsky et al. (1997a), which was developed from the the DAMSL tag set (Core & Allen, 1997). The DAMSL tag set of 220 tags was created using an annotation scheme that classifies and labels utterances from task-oriented dialogues according to their purpose and role in the dialogue. Many of the 220 tags were very infrequently used in the SWBD corpus. The SWBD-DAMSL annotation scheme removed utterances with segmentation or transcription errors. It then clustered and grouped the remaining tags, to remove very small tag classes, resulting in 42 new dialogue act tags. Table 1 shows the 42 resulting SWBD-DAMSL tags, their frequencies in the corpus and an utterance example of each. For example, the most frequent tag *Statement-non-opinion*

(sd) counted for 36.9% of the utterances, with an example being "Me, I'm in the legal department." Figure 2 shows a sample of conversation from the SWDA corpus, where A and B are the two speakers. The dialogue act tags are shown on the left. The slashes represent either the end of an utterance or the end of an interrupted utterance. Potts (2011) reproduced the corpus in an alternative format, which I have used for this study as it was more suitable for the methodology.

In addition to the information as shown in figure 2, the corpus includes part-of-speech (POS) tag and tree information, which Potts added by aligning the corpus with the Penn Treebank3 (Mitchell, Santorini, Marcinkiewicz, & Taylor, 1999). POS tags are a token assigned to each word in a text corpus, based on the definition of the word and its context. These include noun, verb, adjective etc. Potts then created the parse trees automatically, with the help of the POS tags, using a range of heuristic matching techniques. These are rooted trees that represent the syntactic structure of a string of text according to a context-free grammar. A context-free grammar is a set of production rules for generating patterns of strings. A tree includes a non-terminal symbol (root node) and then splits into a number of terminal symbols. These symbols are bracket labels from the Penn Treebank. The bracket labels are sub-divided into three levels - clause (e.g. simple declarative clause (S)), phrase (e.g. noun phrase (NP)) and word level (e.g. past tense verb (VBD)). Figure 1 shows an example from Pott's version of the SWDA corpus, with the root node S (simple declarative phrase).

Figure 1: Parse tree example from Pott's version of the SWDA corpus, with root node S.

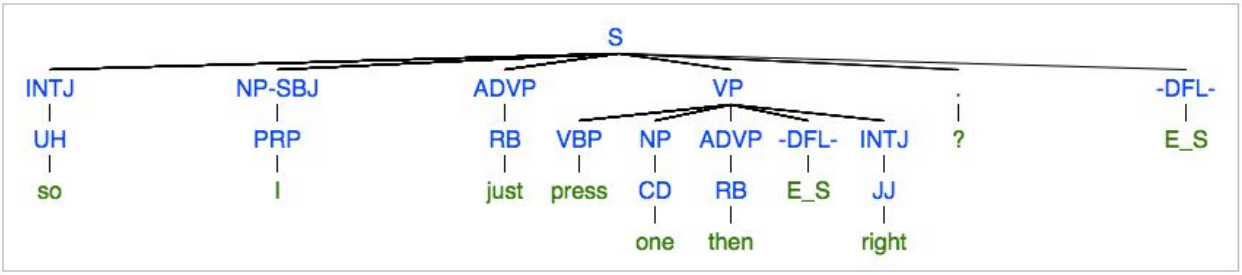

Figure 2: Sample of the SWDA corpus, where A and B are the two speakers. The dialogue act tags are shown on the left.

```
b           A.79 utt1: Yeah,  /
b^r           A.79 utt2: yeah. /

sv           B.80 utt1: That's interesting. /

%           A.81 utt1: {D So, } -/

qy           B.82 utt1: Do you have to have any special training? /

%           A.83 utt1: Not, - /
sd           A.83 utt2: it depends on the state you live actually.  /
sd           A.83 utt3: Some laws absolutely prohibit it.  /
```

I chose the SWDA corpus for a number of reasons:

1. It is a relatively large corpus, so findings from experimentation should hold greater statistical reliability.
2. The corpus is a human-human dialogue corpus, so it reflects natural dialogue interactions and therefore can better enable spoken dialogue systems to operate in an open domain (Serban, Lowe, Henderson, Charlin, & Pineau, 2015; Young, 2015). Serban et al. described natural dialogue as "conversations that are unconstrained and unscripted, e.g. between interlocutors who are not instructed to carry out a particular task, to follow a series of instructions, or to act out a scripted dialogue." They added that human-human dialogue corpora are more diverse than human-machine dialogue corpora and give a dialogue model the opportunity to learn to understand natural language. Apart from being given the topic of conversation, the conversations between the two speakers are wholly natural, with speakers talking as themselves and as they would in a normal conversation, so this corpus should be a good example of a natural language conversation corpus. This is unlike the AMI corpus (Carletta, 2006), which contains conversations between speakers who are playing a role in a fictitious design team and so the conversations are not wholly natural.
3. The corpus is non-task-oriented, which means it can be used to train a dialogue model for use across domains, not just in one specific task domain, such as restaurant bookings (Bordes, Boureau, & Weston, 2017).
4. The corpus has been manually-annotated with dialogue act labels. This was conducted by a large group of linguists in 1997, with an inter-annotator agreement of 0.8 across the group, using the Kappa statistic (Carletta, 1996; Carletta et al., 1997). These labels will act as a gold standard against which to measure the accuracy of the automatic

identification of dialogue acts in this study. The quality of this corpus, including size and labelling, should make it more straightforward to achieve high classification performance when working with classification algorithms for the automatic identification of dialogue act tags later on.

METHODOLOGY

As mentioned, the experimental data is a subset of Jurafsky, Shriberg and Biasca (1997b)'s SWDA corpus, formatted by Potts (2011) with added linguistic information using the Penn Treebank3. Six classifiers were implemented, using the same data and features with the same train-test data split, so that the classifiers would be comparable. Evaluation of the classifier prediction accuracies are detailed below.

Data preprocessing, the machine learning classifiers and evaluation were implemented in Python (3.6.3) with the help of functions from:

- Scikit-learn (v0.19.0)
- Stackoverflow.com (https://stackoverflow.com/questions/43838052/)
- github.com/wendykan/DeepLearningMovies (2014)
- chrisstrelioff.ws (2015)

Data Preprocessing

Removed data

The corpus used in this study has one utterance per data row. These utterances do not all map one-to-one with the parse trees available in the data. Sometimes one tree spans more than one utterance and sometimes one utterance is covered by more than one tree. This makes it more complicated to work with the data, for example when making predictions from the data, which could lead to errors being made. To remove this potential cause of error, I have kept only utterances that have a perfectly-matching tree, meaning that the tree terminals perfectly align to the words in the utterance. Of the original 220,982 utterances, 96,362 remained, so 43% of the data was kept.

I also removed the dialogue act tag '+', which signals segmentation of an utterance with another dialogue act splitting the utterance into two or more parts. This means that one speaker was interrupted by the other speaker and so their utterance was split over two or more data rows, with the rows after the first one having only having the '+' tag. Figure 3 gives an example of this. The '+' tag alone does not indicate any particular dialogue function and because this study looked to classify data rows individually, it was decided to remove all data

with the '+' tag as it would not help the classifiers to receive this input. This only removes 248 rows of data, so should not have a significant effect on the overall prediction results as it is not a significant portion of the data.

Figure 3: An example of a segmented utterance, at the third row, with the '+' dialogue act.

```
qw          A.107 utt2: [ what kind of music [ is, + does ] + # what # --
%           B.108 utt1:  # [ It, + it, ] # -/
+           A.109 utt1:  -- songs does ] he play?  /
sd          B.110 utt1:  [ Th-, + THIS ] LOVE CUTS LIKE A KNIFE.
```

Another modification to the data was to remove some common words that are said to have no great impact in natural language processing, for example "a", "this" and "because". These are known as "*stop words*". There is not yet a definitive list of English stop words, however Scikit-learn has an in-built list of 318 words which can be included in the data preprocessing as a parameter of CountVectorizer and TfidfVectorizer, the vectorizers discussed below, so this study has utilized this list.

For the purpose of simplification in this study, non-letters were also removed using a *regular expression* package. This meant that any characters in a given utterance that were not a lowercase letter (a-z) or uppercase letter (A-Z) were replaced with a space. Stolcke et al. (2000) also removed all punctuation from the corpus for their study and still managed to achieve a high accuracy. The remaining text was then all converted to lowercase, for the benefit of the vectorizers described below.

Features

I also added some features to the data or made modifications to existing ones to allow them to be fed to the classifiers.

Tree root nodes: Tree root node labels were added, for example the figure 1 example would have the tree root node label 'S'. There do not seem to any previous studies specifically looking at parsed tree root nodes as a feature, though some research has used whole trees as features. Ivanovic (2005a) used a probabilistic model with parse

trees to achieve a dialogue act classification accuracy of 84% on a data set of online chat conversations and with dialogue acts taken from the DAMSL tag set. This method is not comparable to this study, however it does show that parse trees are worth investigating with relation to dialogue act classification. Rather than feeding the classifiers with whole trees, I have investigated whether just the less-computationally-expensive root node information can increase dialogue act classification accuracy.

Utterance length: Utterance length was added, as the number of words in the utterance. This length calculation excluded the corpus disfluency markers, as these are not part of the human conversation, they were just added by the annotators during the transcription process. Although a crude feature of prediction, utterance length can be used to make some distinction between dialogue act utterances (Webb, Hepple, & Wilks, 2005; Ferschke, Gurevych, & Chebotar, 2012; Tavafi et al., 2013). For example, Yes answers (ny) and No answers (nn) are usually very short, whereas other dialogue act utterances are usually longer. Webb, Hepple, and Wilks (2005) found utterance length could increase dialogue act classification accuracy by 4%. Hacquebord (2017), on the other hand, found that utterance length was only able to correctly classify 2 of the 39 dialogue act tags included in her study, when it was used as a lone feature.

Bag-of-Words (BoWs): This feature is described in detail below. Grau et al. (2004) achieved a 66% accuracy when using this feature with a naive Bayes classifier fitted to the Switchboard corpus and the DAMSL dialogue act tags. Ivanovic (2005b) also tested the bag-of-words feature with a naive Bayes classifier and DAMSL dialogue act tags, using instant messaging dialogue, however they only used 12 of the tags and they built previous dialogue act probabilities into the model. As a result, they achieved a higher accuracy of 80.6%.

Speaker turn: In the data, the two speakers are labelled as A and B. These were re-coded as 1 and 2, for ease of being fed to the classifiers. Hacquebord (2017) showed speaker turn to increase prediction accuracy, when taking into account who

the previous speaker was. Su, Cavedon, and Baldwin (2010) also found that adding an "author" feature for their online chat corpus increased accuracy.

DAMSL dialogue act tags: The DAMSL tags were added to the corpus, which were a simplification by Jurafsky, Shriberg, and Biasca (1997a) of the already added act tags, as described above. These modifications were made using Python functions from Potts (2011). When splitting data for cross-validation evaluation, there was not enough sample available for all dialogue acts. Therefore, I chose to focus more on the top 10 most frequent dialogue acts. As the Switchboard corpus is domain-independent and not task-oriented, these should be more reflective of which dialogue acts are most prevalent in general human conversations and therefore also most worth investigating for application in domain-independent and non-task-oriented conversation agents. As mentioned in the *Classifiers in previous studies* section, focusing on fewer dialogue act tags in classification can lead to an increased accuracy score, therefore I expect my results to reflect this, when compared to other studies such as Stolcke et al. (2000).

Text cannot usually be fed directly to ML classifiers, as they usually expect numerical vector input. Therefore, a process of text feature extraction - creating a "bag-of-words" representation - was carried out on the utterance text data in this study. To prepare the data for this process, the utterances are taken from the data and "cleaned". All symbols except letters of the alphabet were removed from the utterances. Each utterance, i.e. the text in each row of the data, is represented as a set of its words. The words are tokenized, meaning that the utterance is broken down into a string of its words, or "tokens". The amount of times that the tokens appear in the document, or utterance, are counted. This is done for all sample utterances.

In this study, two of Scikit-learn's vectorizers, "CountVectorizer" and "TfidfVectorizer", were implemented and used for the tokenization and counting. CountVectorizer orders the tokens by occurrence count and indexes them according to this order. This is used to make an array for the bag of words, with each column representing one token or "feature" and each row representing a row from the input data as a vector. With a large dataset with a wide

vocabulary, there will be a very large number of features, therefore a max_features parameter was specified, as there will be many features that add to processing time but do not give important information for the classification process. This was set at 500, so the array row was a vector of size 500. There are 25,495 unique utterance tokens in the modified data (a crude value including punctuation and other markers), so a much higher max_features value could have been considered, however due to the large data size, 500 was chosen so that 'runtime' was not too long during experimentation. As the vectorizer has ordered the tokens according to occurrence count, the first number in the vector represents how often the most common token in the dataset occurs in the utterance corresponding to that row of the array. An utterance will only be a very small subset of the total number of words in the whole dataset, for example the average number of words in an utterance in this study is 17, so the bag of words arrays produced by the vectorizers will mostly consist of zeros. With text classification tasks, this will usually be more than 99% of the feature values (Scikit-learn, 2011).

TfidfVectorizer runs the same process as CountVectorizer, except with the addition of applying normalization the the matrix of occurrence counts using "Term Frequency Inverse Document Frequency" (TF-IDF) normalization. This is a function commonly used in information retrieval and text mining. The idea is to remove the impact of high-frequency words in the training data, as these will often be words such as 'a' and 'the', which have very little importance in improving the classification process. The process calculates how "unique" a word is to a particular document within its corpus, using the equation TF x IDF. This is the word frequency multiplied by the inverse document frequency, which is a measure of how much information the word provides in the corpus. A word with a high tf-idf value will occur frequently in a document, or utterance, but infrequently across the other documents in the corpus.

Data Partitioning

The data was split into a training set, consisting of 75% of the data (72,278 samples), and a test set, consisting of 25% of the data (24,092 samples). The class labels were also split from the data, so each set was split into the unlabelled data and the class labels. Each classifier in turn was then fitted to the unlabelled training data and labels, to train the classifiers. For all classifiers, an integer was specified for the random_state parameter, so that the classifiers

could be better compared. This split was used for calculating the precision, recall and F1-scores, as described below. For cross-validation evaluation, described below, the cross-validation process ran its own data splitting.

Evaluation of Classifiers

Cross-validation was used in the classifier evaluation process to reduce any possible overfitting. Overfitting is when parameters are tuned based on the training data, so that the model fits the training data well, but when the test data is then fed to the model, the model does not fit this data as well. Cross-validation is a process that ensures that, as much as possible, the criteria used for training the model are the same as those used to evaluate the model in the test phase. To achieve this, the data is partitioned into training and test data and the model is evaluated with cross-validation. This is repeated for multiple other rounds of cross-validation and the results of the multiple rounds are analysed together, for example averaged, to give a final prediction accuracy for the model. In this study, I used 5- and 10-fold cross-validation. This means that the data went through the cross-validation process two times - once with 5 rounds and once with 10 rounds of cross-validation. These are separate of each other. I chose to use 5- and 10-folds as these are commonly used for classifier evaluation.

The accuracy of the classifiers was also evaluated by calculating their precision, recall and F1-score when making dialogue act predictions, to give an alternative form of evaluation to cross-validation. The Scikit-learn function used for this also outputs a "support" value, which gives the number of occurrences of each of the target labels, i.e. the each of the 42 dialogue tags, in this study. The three metrics - precision, recall and F1-score - can be outputted for each dialogue act and they can also be outputted as the weighted averages of all dialogue acts, weighted using the support values. In this study, both metric outputs were investigated.

Precision: this measures the ability of the classifier to not label a negative sample as positive, i.e. it measures how many of the utterances were classified with the correct dialogue act tag. This is calculated with the equation:

PRECISION = TRUE POSITIVES / (TRUE POSITIVES+FALSE POSITIVES)

Recall: this measures the ability of the classifier to find all of the positive samples, i.e. it measures how many utterances of a certain dialogue act tag were classified with the correct tag. This is calculated with the equation:

RECALL = TRUE POSITIVES / (TRUE POSITIVES+NEGATIVES)

F1-score: also known as F-measure; a harmonic mean between the precision and recall, ranging between 0 and 1. This is calculated with the equation:

F1-SCORE = 2 x (PRECISION x RECALL) / (PRECISION + RECALL)

RESULTS and ANALYSIS

After the data modifications described in the *Methodology* section, the features (speaker turn, utterance length and root node) with both the CountVectorizer and TfidfVectorizer bag-of-words features were fitted by the 6 classifiers described in the *Machine Learning Classifiers* section. Due to the large size of the dataset, the SVM selected was a linear SVM, as opposed to the C-support Vector classifier, as the C-support Vector classifier was running with a very long fitting time. As described by Scikit-learn, this is because the "fit time complexity is more than quadratic with the number of samples which makes it hard to scale to dataset with more than a couple of 10000 samples." One baseline classifier was compared to five other classifiers. Once the data was preprocessed and some data removed, the baseline changed from 36.9% to 39%. This was because the most frequent dialogue act tag was now assigned to 39% of the utterances. So, the five classifiers were tested against a baseline of 39%. The features were also tested individually, to identify their individual impact on the prediction accuracies. These are discussed below. Two methods of evaluation were used - the Scikit-learn metrics algorithm *precision_recall_fscore_support* and cross-validation.

The data modifications changed the dialogue act frequency counts slightly, from the original corpus, though the top 3 acts remain the same and still make up the majority of the corpus. These are *statement-non-opinion(sd)(38.5%)*, *acknowledge-backchannel(b)(22.4%)*, and *statement-opinion(sv)(12.7%)*, with the full top 10 shown in Table 2 and the percentages representing proportion of the top 10 dialogue acts. As discussed in the *Methodology* section, I have focused on these top 10 dialogue acts in further analysis.

Precision, Recall and F1-score Results

Precision_recall_fscore_support was used to evaluate the accuracy of these classifiers and the results can be found in Table 3 and Table 4. Table 3 are results from evaluating results from the top 10 most frequent dialogue acts, where the classifiers were fitted only to data for the top 10 dialogue acts. Table 4 are the results from evaluating all dialogue act results, where the classifiers were fitted to data from all dialogue acts. Comparing the two tables, the effect of focusing on just the top 10 dialogue acts, as opposed to all 42 acts, can be seen. As the baseline classifier simply predicts all dialogue acts to have the most frequent dialogue act

label, it is not affected by which vectorizer is used, so the CountVectorizer and TfidfVectorizer have the same baseline average results.

When evaluating data of all dialogue acts, the multilayer perceptron classifier had the highest precision, recall and F1-score accuracies for both the CountVectorizer and TfidfVectorizer. These were highest for the TfidfVectorizer, with precision=63.3%, recall=67.4% and F1-score=63.0%. The linear support vector machine classifier with the TfidfVectorizer had the lowest accuracies for all dialogue acts (precision=39.5%, recall=51.2% and F1-score=41.4%). These are compared to the baseline average for precision (11.7%), recall (34.2%) and F1-score (17.4%) for all dialogue acts.

When focusing on evaluating data of just the top 10 dialogue acts, the random forest classifier had the highest precision, recall and F1-score accuracies for the CountVectorizer (66.6%, 69.0% and 65.9%, respectively). For the TfidfVectorizer, the random forest had the highest recall (70.0%) but the multilayer perceptron had the highest precision (68.3%) and F1-score (66.7%). These are compared to the baseline average for precision (14.2%), recall (37.7%) and F1-score (20.6%) for the top 10 acts. The naive Bayes classifier with the TfidfVectorizer had the lowest accuracies for the top 10 dialogue acts (precision=61.1%, recall=64.1% and F1-score=57.2%).

It is not clear from these results if the increased accuracy, when only training the models and making predictions for the top 10 dialogue acts, is due either the larger amount of training data available for these 10 acts, the increased ease of classification of these acts, or a combination of both.

Cross-validation Results

Cross-validation was run on only the top 10 most frequent dialogue acts because, in some cases, less frequent dialogue acts did not have enough sample to be represented in each fold of cross-validation. The mean cross-validation score and the 95% confidence interval of the score estimate were calculated for both 5-fold and 10-fold cross-validation. The mean accuracy score of each of these was calculated with a 95% confidence interval, meaning that we can be 95% certain that the true accuracy score lies within this interval range. These

results are shown in Table 5. For example, with 10-fold cross-validation the decision tree classifier achieved 70% accuracy with a 95% confidence interval of +/- 2%. This means there is 95% chance that the true accuracy is within 2% of 70% accuracy.

Each classifier achieved the same accuracies for both 5- and 10-fold cross-validation and both the CountVectorizer and TfidfVectorizer, except the naive Bayes classifier, which had a 1% higher accuracy for the 10-fold cross-validation with the TfidfVectorizer. 5-fold cross-validation generally showed narrower confidence intervals. The random forest and the multilayer perceptron classifiers both achieved the highest accuracy of 71%. The random forest classifier achieved this for both vectorizers, whilst the multilayer perceptron classifier achieved 71% for the TfidfVectorizer and 70% for the CountVectorizer. The decision tree classifier achieved the next highest accuracies (70%), followed by the linear support vector machine classifier (68 to 69%) and then the naive Bayes classifier (66 to 67%).

Individual Features

The individual features were tested using the data for the top 10 dialogue acts.

Tree root nodes: Table 6 shows the top 10 tree root nodes for all utterances from the modified study data. 89.5% of the utterances have either a simple declarative clause (S) root node (55.9%) or an interjection (INTJ) root node (33.6%). A potential issue with the tree root nodes in this corpus data is that they include tags from different levels of the Penn Treebank tagset, for example INTJ is from the phrase level but can also be a node of a simple declarative clause. When the classifiers had only the root nodes feature to fit to, the average classifier accuracies were relatively high. The decision tree and random forest classifiers reached F1-scores of 53.8% and 53.9%, respectively, and both achieved mean accuracy scores of 67% (+/- 1%) for 5- and 10-fold cross-validation. These scores are all a significant increase on the baseline accuracy of 39%.

Utterance length: With utterance length as the only feature to fit to, the classifiers performed surprisingly well. The decision tree, random forest and multilayer perceptron classifiers all achieved mean accuracy scores of 60% (+/- 0%) for 5-fold cross-validation and 60% (+/- 1%)

for 10-fold cross-validation. These three also achieved F1-scores of 47.4%. Both scores are a significant increase on the baseline accuracy of 39%.

Speaker turn: With speaker turn as the only feature to fit to, the classifiers only performed as well as baseline (39%) and the F1-scores were only 20.6%. This is not surprising, as the speakers were not assigned different roles in the conversations and so it would not be expected that the two speakers would use differing dialogue acts. Therefore, it is difficult to predict a dialogue act based on which speaker an utterance belonged to.

Bag-of-words: With bag-of-words as the only feature to fit to, the linear SVM achieved the highest accuracies (64% (+/-1 and 2%) for cross-validation and CountVectorizer F1-score of 57.7%). This is in contrast to the results for all features combined, where the linear SVM linear classifier was one of the lowest-performing classifiers. There were not large differences between the classifier performances, however. The decision tree classifier achieved the lowest cross-validation mean accuracy of 62% (+/-2%). Again, these scores are all a significant increase on the baseline accuracy of 39%.

Cross-validation evaluation was also run on the classifiers when adding the features one-by-one for the top 10 most frequent dialogue acts. As mentioned previously, the results were the same in 5- and 10-fold cross-validation and for both vectorizers, apart from a 1% difference with the naive Bayes classifier, so I have shown the results for 10-fold and the TfidfVectorizer in Table 7. The most significant finding here is that adding root node as a feature along with bag-of-words increased the accuracy by 7% for the decision tree, random forest and MLP classifiers. Utterance length and speaker turn then led to little or no increase in accuracy.

DISCUSSION

Key Findings

In this study, a dialogue act classification accuracy of 71% was reached by the random forest and multilayer perceptron classifiers, when averaging cross-validation scores for the top 10 most frequent dialogue acts. The multilayer perceptron classifier also had the highest precision (63.3%), recall (67.4%) and F1-score (63.0%) when evaluating the data for all 42 dialogue acts, whilst the random forest, decision tree and multilayer perceptron classifiers shared the highest accuracy scores amongst them for these accuracy measures with the top 10 most frequent dialogue acts. The linear support vector machine and naive Bayes classifiers performed least well in both precision, recall and F1-scores and cross-validation, though there was not a large difference in performance between these and the top-performing classifiers.

Individually, the tree root node, utterance length and bag-of-words features used in this study yielded accuracies of 67%, 60% and 64%. Speaker turn, as expected, yielded no increase in accuracy beyond baseline when used alone. When adding features one-by-one, beginning with bag-of-words, adding the root node feature with bag-of-words increased accuracy by 7% for 3 of the 5 classifiers.

Relation of Findings to Study Aims

I successfully fulfilled my first aim by fitting five machine learning classifiers, plus a baseline classifier, to four dialogue text features. This led to the automatic identification of dialogue acts with average accuracies of up to 71%. The second aim was to explore parse tree root nodes as a feature for classification. I successfully began this exploration, showing it being a worthwhile feature as, individually, this yielded a 67% mean accuracy score. Aside from ‘speaker turn’ as a lone feature, all feature combinations yielded accuracies in the upper range of the predicted range of between 39% (baseline) and 84% (gold standard).

Relation of Findings to Previous Research

This study used a transcribed conversation corpus, dialogue act tags and machine learning classifiers that have been used in previous studies of dialogue act classification. It also covered a couple of areas where there is no or little literature, namely random forest classifiers in dialogue act classification and the use of parse tree root nodes as a feature. This

study achieved similar mean accuracies to similar previous studies. For example, Stolcke et al. (2000) used the same corpus and DAMSL tag set and achieved a 71% accuracy, which is in line with our F1-score of 63% for the whole tag set and 71% cross-validation accuracy for the top 10 most frequent tags. Tran, Zukerman and Haffari (2017) also achieved a relatively high accuracy (74.5%) with the Switchboard corpus, compared to their accuracy achieved on another corpus, and Khanpour, Guntakandla and Nielsen (2016) managed to achieve a yet higher accuracy (80.1%), with the Switchboard corpus and DAMSL tag set, by training an RNN with long-distance dependencies.

Previous studies described above had pointed to the beneficial effect of focusing on just a selection of a dialogue act tag set (e.g. Verbree, Rienks, & Heylen, 2006; Tavafi et al., 2013). Findings from this study also supported this idea, with an increase in accuracy scores when focusing only on the top 10 most frequent dialogue act tags.

The results of classifier comparison studies were mixed, in previous studies. However, this study supported the strength of neural networks (e.g. Khanpour, Guntakandla, & Nielsen, 2016), with results from a multilayer perceptron, and it also supported the findings from previous studies (e.g. Moldovan, Rus, & Graesser, 2011); Samei et al., 2014) that showed naive Bayes classifiers to be outperformed by other classifiers.

Unexpected Results

In general, the findings were in line with findings from previous studies. One finding that did stand out, though, was related to utterance length. Considering that utterance length, as a lone feature, was not expected to have a great effect on classification accuracy, it achieved a relatively high mean accuracy of 60% with the top 10 most frequent dialogue act tags. Perhaps limiting the number of tags makes this feature more effective, for example Tavafi et al. (2013) used a limited number of dialogue act tags and found utterance length to be a "very effective" feature.

Significance of Findings

Of key significance is the findings that parse tree root node information added 7% to the accuracy score, when added as a feature along with bag of words, and achieved an accuracy

of 67% when just used as a lone feature in classification. In real-world application, such as with conversational agents, full parse trees are already being used. However, these root node findings suggest that simple classification decisions, at least, could be more quickly made by just looking at the root nodes. Another significance of this study is that there had not been much previous literature on random forest and multilayer perceptrons in dialogue act classification, so this study is providing new support for the use of these.

Limitations and Improvements for Future Work

In this study, I have investigated four features - speaker turn, utterance length, root node and bag-of-words. These have successfully contributed to achieving a certain degree of accuracy in the automatic identification of dialogue acts, however there are additional features that other previous studies have used to increase accuracy and these could help improve the accuracy in future work related to this study. Many studies use n-grams as a feature (e.g. Stolcke et al., 1998; Surendran & Levow, 2006; Ribeiro et al., 2015). An n-gram is a sequence of items, such as words or part-of-speech (POS) tags, from a certain sequence of text or speech. An n-gram of one item is called a unigram, a two item n-gram is called a bigram, a three item n-gram is called a trigram, etc. They allow predictions to be made based on sequences of dialogue acts, for example a trigram of the previous three dialogue acts. N-grams could be used to improve the accuracies of this study.

A limitation of my study that will apply to dialogue act classification studies in general, is that some dialogue acts had a very low amount of data. Hacquebord (2017) showed that this can affect classification accuracy. She evaluated classification accuracy by each dialogue act and found that dialogue acts with fewer than 100 utterances in the test set were identified with less accuracy by a classifier than those with above 100 utterances. Lack of data for certain dialogue acts also affected my study, as I was not able to run cross-validation for all dialogue acts because there was insufficient data for them. I had to run cross-validation on only the top 10 dialogue acts due to this.

Dialogue corpora are often annotated with the "gold standard" human labelling using inter-annotator agreement, as mentioned earlier. There are a couple of limitations here for my study and other dialogue act classification studies that use this gold standard. Firstly, there

was only 84% inter-annotator agreement, despite there being an extensive manual available to the annotators who were experts in the field. As a result, we cannot be sure that our achieved accuracy is the true accuracy, as the true dialogue act tags are not 100% agreed upon. Secondly, as mentioned before, there is not yet an agreed standard for dialogue acts as these so far depend on human intuition and differ according to corpus domain and whether or not it is task-oriented. This makes it difficult for annotators to be consistent and agree on dialogue act tags. Rus et al. (2012) experimented with an alternative to human labelling, which may help lead to truer labelling. They used clustering algorithms with features from an education corpus to identify speech act categories. A human then gave a semantic label to the categories. Their methodology and findings still carry the limitation of being domain specific and dependent on the features that they selected, however the methodology could be built on to be more generalisable.

There is some metadata available that could be used to improve classification in this study. Potts (2011) combined available metadata for the Switchboard corpus with the Switchboard Dialogue Act Corpus, which I did not use for the purpose of this study. However, some of this information could be used to create features to fit the classifiers to. These are conversation-level as opposed to utterance-level, for example topic descriptions and prompts such as “VACATION SPOTS” and “PLEASE DISCUSS TYPES OF VACATIONS AND TRIPS YOU ENJOY. FIND OUT WHETHER THE OTHER CALLER CAN INTEREST YOU IN A VACATION SPOT YOU HAVEN'T VISITED.”, respectively. These do not appear to have been used yet in dialogue act classification research with the Switchboard corpus. Some topics may elicit more emotive dialogue acts or have shorter utterance lengths, for example, which a classifier could learn. This would effectively mean training the classifier with domain-specific corpus data across multiple domains, as each topic is a domain, though this may actually require training a separate classifier for each topic.

In this study, I very crudely removed all non-letters from the utterances. This included removing disfluency markers, which had been added when the corpus was built, punctuation and numerical numbers. Some of the lexical features here could be used to improve on the accuracies achieved in my study, for example a question mark would clearly be indicative of one of the question dialogue acts. Ribeiro et al. (2015) used punctuation as a feature in

dialogue act classification and suggested that it may help disambiguate differing intentions related to the same words, for example an exclamation may change a statement to a command. Hacquebord (2017) found, however, that having a question mark as a feature did not approve classification accuracy by much.

This study focused on written transcript corpora, however there is also audio data available for the Switchboard corpus. Although Shriberg et al. (1998) showed that adding acoustic features added little improvement to the classification accuracy, many studies have used acoustic features to increase classification accuracy (e.g. Stolcke et al., 2000) and so this looks to be worth exploring.

CONCLUSION

Five machine learning classifiers were fitted to four text features from a modified version of the human-human transcribed conversation Switchboard Dialogue Act corpus and used to automatically predict dialogue acts. Accuracy was evaluated using precision, recall and F1-scores and cross-validation, with the "true" dialogue act labels coming from expert corpus annotators. Focus was placed on the top ten most frequent dialogue acts and, with these, the random forest and multilayer perceptron classifiers achieved average accuracies of 71%. A feature of note was the parse tree root nodes feature. This helped increase accuracy by 7%.

Future work can continue to work towards a higher level of accuracy, for example by including acoustic data and other features of the SWDA corpus. Once there is an agreed standard for dialogue act annotation, potentially via a combination of clustering algorithms and human labelling, there can be more progress made in automatic dialogue act classification.

REFERENCES

Allen, J., & Core, M. (1997). DAMSL: Dialog act markup in several layers. *Technical Report draft 2.1, Multiparty Discourse Group, Discourse Research Initiative.*

AMI Consortium (2005). Guidelines for Dialogue Act and Addressee Annotation. *Technical Report.*

Andernach, T., Poel, M. & Salomons, E. (1997). Finding classes of dialogue utterances with kohonen networks. *Workshop Notes of the ECML / MLnet Workshop on Empirical Learning of Natural Language Processing Tasks*, 85–94.

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., ... Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, *34,* 351-366.

Bordes, A., Boureau, Y., & Weston, J. (2016). Learning End-to-End Goal-Oriented Dialog. Facebook AI Research. *arXiv:1605.07683v4.*

AQA.org.uk. (2015). Glossary of key terms and guide to methods of language analysis [PDF document]. Retrieved from http://filestore.aqa.org.uk/resources/english/AQA-7706-7707-GLOSSARY-CTT.PDF

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. *Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.*

Breiman, L. (2001). Random Forests. *Machine Learning, 45(1),* 5-32.

Bunt, H., Alexandersson, J., Choe, J., Chengyu Fang, A., Hasida, K., Petukhova, V. ... Traum, D. (2012). ISO 24617-2: A semantically-based standard for dialogue annotation. *In Proceedings of 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul. Paris: ELDA.*

Burger, S., Weilhammer, K., Schiel, F. & Tillmann, H. (2000). Verbmobil data collection and annotation. *Verbmobil: Foundations of speech-to-speech translation,* 537–549.

Byron, D., & Stent, A. (1998). A preliminary model of centering in dialog. *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Quebec, Canada,* 1475-7.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics 22,* 249-254.

Carletta, J. (2006). Announcing the AMI Meeting Corpus. *The ELRA Newsletter 11(1), January-March,* 3-5.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1997). The Reliability of A Dialogue Structure Coding Scheme. *Computational Linguistics 23.1,* 13-32.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1996). *HCRC Dialogue Structure Coding Manual (HCRC/TR-82). Edinburgh, Scotland: Human Communication Research Centre, University of Edinburgh.*

Dbabis, S., Ghorbel H., Belguith, L., & Kallel, M. (2015). Automatic Dialogue Act Annotation within Arabic Debates. Gelbukh A. *(eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science, vol 9041.*

Eckert, M. & Strube, M. (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics, 17:51–89.*

Ferschke, O., Gurevych, I., & Chebotar, Y. (2012). Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. *Proceedings of the 13th Conference of the European Chapter of the ACL.*

Forsythand, E., & Martell, C. (2007). Lexical and Discourse Analysis of Online Chat Dialog. *International Conference on Semantic Computing (ICSC 2007). IEEE,* 19–26.

John, G., & Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62. *Philadelphia: Linguistic Data Consortium.*

Grau, S., Sanchis, E., Castro, M., & Vilar, D. (2004). Dialogue Act Classification Using a Bayesian Approach. *Proceedings of the 9th International Conference Speech and Computer,* 495–499.

Hacquebord, L. (2017). Dialogue Act Recognition for Conversational Agents. *Master Thesis, Utrecht University.*

Ho, T. (1995). Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC,* 278–282.

Gang, J., & Bilmes, J. (2005). Dialog Act Tagging Using Graphical Models. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing. Vol. 1. IEEE,* 33–36.

Hu, J., Passonneau, R., & Rambow, O. (2009). Contrasting the interaction structure of an email and a telephone corpus: A machine learning approach to annotation of dialogue function units. *Proceedings of the SIGDIAL 2009 Conference, London, UK*.

Ivanovic, E. (2005a). Automatic utterance segmentation in instant messaging dialogue. *Proceeding of The Australasian Language Technology Workshop,* 241-249.

Ivanovic, E. (2005b). Dialogue Act Tagging for Instant Messaging Chat Sessions. *Proceedings of the ACL Student Research Workshop,* 79–84.

Janin, A. et al., 2003. The ICSI Meeting Corpus. Proc. ICASSP- 2003.

Jekat, Susanne, Alexandra Klein, Elisabeth Maler, Ilona Maleck, Marion Mast, and J. Joachim Quantz. 1995. Dialogue Acts in VERBMOBIL. Verbmobil Report 65, Universitiit Hamburg, DFKI Saarbriicken, Universitiit Erlangen, TU Berlin.

Minwoo Jeong, Chin-Yew Lin, and Gary G. Lee. 2009. The Semi-supervised speech act recognition in emails and forums. Proceedings of the 2009 Conf. Empirical Methods in Natural Language Processing.

Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca. 1997a. Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, Draft 13, University of Colorado, Boulder. Institute of Cognitive Science Technical Report 97-02.

Jurafsky, Daniel, Elizabeth Shriberg, and Debra Biasca. 1997b. Switchboard Dialog Act Corpus. Corpus can be downloaded at http://web.stanford.edu/~jurafsky/ws97/

Kan, Wendy, (2014), Deep Learning Movies, GitHub repository, https://github.com/wendykan/DeepLearningMovies

Khanpour, H.; Guntakandla, N.; and Nielsen, R. 2016. Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In COLING

Kim, Su Nam, Lawrence Cavedon, and Timothy Baldwin (2010). “Classifying Dialogue Acts in One-on-one Live Chats”. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP ’10)*. Cambridge, Massachusetts: Association for Computational Linguistics, pp. 862–871.

Kral, P., & Cerisara, C. (2014). Automatic dialogue act recognition with syntactic features. *Language Resources and Evaluation 48(3),* 419–441.

Mitchell, M., Santorini, B., Marcinkiewicz, M., & Taylor, A. (1999). Treebank-3 LDC99T42. *Web Download. Philadelphia: Linguistic Data Consortium.*

Moldovan, C., Rus, V., & Graesser, A. (2011). Automated Speech Act Classification For Online Chat. *22nd Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2011). MAICS,* 23– 29.

Moore, J., & Wiemer-Hastings, P. (2003). Discourse in computational linguistics and artificial intelligence. *A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), Handbook of Discourse Processes,* 439–485. *Mahwah, NJ: Erlbaum.*

O'Shea, J., Bandar, Z., & Crockett, K. (2012). A Multi-Classifier Approach to Dialogue Act Classification Using Function Words. *Lecture notes in computer science, 7270,* 119-143.

Passonneau, R., & Sachar, E. (2014). Loqui Human-Human Dialogue Corpus. *(Transcriptions and Annotations), Columbia University Academic Commons,* doi.org/10.7916/D82R3PW9.

Potts, C. (2011). The Switchboard Dialog Act Corpus. *LING7800-007: Computational Pragmatics. LSA Linguistic Institute 2011: Language in the World. Stanford Linguistics.* Retrieved from http://compprag.christopherpotts.net/swda.html

Pranckevicius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing, 5, 2,* 221-232*,* doi.org/10.22364/bjmc.2017.5.2.05.

Ribeiro, E., Ribeiro, R., & Martins de Matos, D. (2015). The influence of context on dialogue act recognition. *arXiv preprint arXiv:1506.00839.*

Rus, V., Moldovan, C., Niraula, N., & Graesser, A. (2012). Automated Discovery of Speech Act Categories in Educational Games. *Proceedings of International Conference on Educational Data Mining,* 25-32.

Samei, B., Li, H., Keshtkar, F., Rus, V., & Graesser, A. (2014). Context-Based Speech Act Classification in Intelligent Tutoring Systems. *In: Trausan-Matu, S., Boyer, K., Crosby M., & Panourgia, K. (eds). Proceedings of the 12th International Conference on Intelligent Tutoring Systems,* 236–241.

Schonfeld, E. (2010, February 4). Siri's IPhone App Puts A Personal Assistant In Your Pocket. *Retrieved from https://techcrunch.com/2010/02/04/siri-iphone-personal-assistant/.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Searle, J. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Searle, J. (1976). A classification of illocutionary acts. *Language in society 5.1,* 1–23.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., … Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech, 41, 3-4,* 439-487.

Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., & Carvey, H. (2004). The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Proceedings of 5th SIGdial Workshop on Discourse and Dialogue,* 97-100.

Loos, E., Anderson, S., Day, D., Jordan, P., & Wingate, J. (2003). Glossary of Linguistic Terms. Retrieved from http://www.glossary.sil.org/.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., & Van Ess-Dykema, C. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics, 26, 3,* 339-371.

Serban, I., Lowe, R., Henderson, P., Charlin, L., & Pineau, J. (2015). A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *arXiv preprint arXiv:1512.05742.*

Strelioff, C. (2015). Decision trees in python with scikit-learn and pandas. Retrieved from http://chrisstrelioff.ws/

Kim, S., Cavedon, L., & Baldwin, T. (2010). Classifying dialogue acts in one-on-one live chats. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing,* 862–871.

Surendran, D. & Levow, G. (2006). Dialog act tagging with support vector machines and hidden markov models. *Interspeech 2006 – Proceedings of the International Conference on Spoken Language Processing,* 1950–1953.

Tavafi, M., Mehdad, Y., Joty, S., Carenini, G., & Ng, R. (2013). Dialogue act recognition in synchronous and asynchronous conversations. *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2013), 13. Association for Computational Linguistics,* 117–121.

Tran, Q., Zukerman, I., & Haffari, G. (2017). A hierarchical neural model for learning sequences of dialogue acts. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 1,* 428–437.

Traum, D., & Hinkelman, E. (1992). Conversation acts in task-oriented spoken dialogue. *Computational Intelligence, 8, 3,* 575–599.

Verbree, D., Rienks, R., & Heylen, D. (2006). Dialogue-act tagging using smart feature selection; results on multiple corpora. *2006 IEEE Spoken Language Technology Workshop*, 70–73.

Walker, M., & Whittaker, S. (1990). Mixed initiative in dialogue: An investigation into discourse segmentation. *Proceedings of the 28th Annual Meeting, Association for Computational Linguistics,* 70-78.

Webb, N., Hepple, M., & Wilks, Y. (2005). Dialogue act classification based on intra-utterance features. *Proceedings of the AAAI Workshop on Spoken Language Understanding*.

Wright, H. (2000). Modelling Prosodic Information for Automatic and Dialogue Speech Recognition. *Ph.D. thesis, University of Edinburgh.*

Wu, T., Khan, F., Fisher, T., Shuler, L., & Pottenger, W. (2005). Posting Act Tagging Using Transformation-Based Learning. *Foundations of Data Mining and knowledge Discovery. Ed. by Tsau Young Lin et al.,* 319–331.

Young, S. (2015). Open Domain Statistical Spoken Dialogue Systems [PDF document]. Retrieved from http://grammars.grlmc.com/SLSP2015/Slides/d3s1ODSDS.pdf

APPENDIX

Table 1: Counts of the 42 tags from the SWDA corpus and utterance examples of each tag.

	SWBD-DAMSL Tag	Tag Label	Examples	Total Count	% Count
1	sd	Statement-non-opinion	Me, I'm in the legal department.	75143	36.9
2	b	Acknowledge (Backchannel)	Uh-huh.	38298	18.8
3	sv	Statement-opinion	I think it's great	26426	13.0
4	%	Abandoned or Turn-Exit; Uninterpretable	But, uh, yeah	15550	7.6
5	aa	Agree/Accept	That's exactly it.	11133	5.5
6	ba	Appreciation	I can imagine.	4764	2.3
7	qy	Yes-No-Question	Do you have to have any special training?	4726	2.3
8	x	Non-verbal	[Laughter], [Throat_clearing]	3630	1.8
9	ny	Yes answers	Yes.	3034	1.5
10	fc	Conventional-closing	Well, it's been nice talking to you.	2582	1.3
11	qw	Wh-Question	Well, how old are you?	1979	1.0
12	nn	No answers	No.	1377	0.7
13	bk	Response Acknowledgement	Oh, okay.	1306	0.6
14	h	Hedge	I don't know if I'm making any sense or not.	1226	0.6
15	qy^d	Declarative Yes-No-Question	So you can afford to get a house?	1218	0.6
16	bh	Backchannel in question form	Is that right?	1053	0.5
17	^q	Quotation	You can't be pregnant and have cats	983	0.5

18	bf	Summarize/reformulate	Oh, you mean you switched schools for the kids.	952	0.5
19	fo_o_fw_"_by_bc	Other	Well give me a break, you know.	883	0.4
20	na	Affirmative non-yes answers	It is.	847	0.4
21	ad	Action-directive	Why don't you go first	746	0.4
22	^2	Collaborative Completion	Who aren't contributing.	723	0.4
23	b^m	Repeat-phrase	Oh, fajitas	688	0.3
24	qo	Open-Question	How about you?	656	0.3
25	qh	Rhetorical-Questions	Who would steal a newspaper?	575	0.3
26	^h	Hold before answer/agreement	I'm drawing a blank.	556	0.3
27	ar	Reject	Well, no	345	0.2
28	ng	Negative non-no answers	Uh, not a whole lot.	302	0.1
29	br	Signal-non-understanding	Excuse me?	298	0.1
30	no	Other answers	I don't know	285	0.1
31	fp	Conventional-opening	How are you?	225	0.1
32	qrr	Or-Clause	or is it more of a company?	209	0.1
33	arp_nd	Dispreferred answers	Well, not so much that.	207	0.1
34	t3	3rd-party-talk	My goodness, Diane, get down from there.	117	0.1
35	oo_co_cc	Offers, Options, Commits	I'll have to check that out	110	0.1
36	aap_am	Maybe/Accept-part	Something like that	105	0.1
37	t1	Self-talk	What's the word I'm looking for	103	0.1
38	bd	Downplayer	That's all right.	103	0.1
39	^g	Tag-Question	Right?	92	0.0
40	qw^d	Declarative Wh-Question	You are what kind of buff?	80	0.0

41	fa	Apology	I'm sorry.	79	0.0
42	ft	Thanking	Hey thanks a lot	78	0.0

Table 2: Frequency of the DAMSL act tags in the study data.

DAMSL act tag	Frequency in data	% Frequency
sd	33623	38.5%
b	19519	22.4%
sv	11129	12.7%
%	8101	9.3%
aa	6232	7.1%
ba	2565	2.9%
qy	2323	2.7%
ny	1680	1.9%
fc	1344	1.5%
bk	814	0.9%

Table 3: Accuracy scores for the 6 classifiers, split by the two bag-of-words methods, for the top 10 most frequent DAs

	CountVectorizer			TfidfVectorizer		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Majority (most freq.)	14.2%	37.7%	20.6%	14.2%	37.7%	20.6%
Decision Tree	64.6%	67.3%	64.3%	66.8%	69.2%	66.1%
Random Forest	66.6%	69.0%	65.9%	67.6%	70.0%	66.5%
Multilayer Perceptron	66.1%	67.3%	65.0%	68.3%	69.7%	66.7%

Naive Bayes	61.5%	64.9%	59.6%	61.1%	64.1%	57.2%
Linear SVM	61.8%	65.3%	62.1%	62.3%	66.9%	62.7%

Table 4: Accuracy scores for the 6 classifiers, split by the two bag-of-words methods, for all DAs

	CountVectorizer			**TfidfVectorizer**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Majority (most freq.)	11.7%	34.2%	17.4%	11.7%	34.2%	17.4%
Decision Tree	51.3%	46.1%	45.8%	59.7%	62.7%	60.7%
Random Forest	59.7%	56.5%	55.4%	61.9%	64.9%	60.5%
Multilayer Perceptron	61.2%	60.4%	58.6%	63.3%	67.4%	63.0%
Naive Bayes	55.8%	60.1%	56.0%	52.1%	58.1%	51.8%
Linear SVM	40.1%	51.5%	42.1%	39.5%	51.2%	41.4%

Table 5: Five- and ten-fold cross-validation mean results for top 10 dialogue acts, with the 95% confidence intervals given in parentheses.

	CountVectorizer		**TfidfVectorizer**	
	5-fold	**10-fold**	**5-fold**	**10-fold**
Majority (most freq.)	0.39 (+/- 0.00)	0.39 (+/- 0.00)	0.39 (+/- 0.00)	0.39 (+/- 0.00)
Decision Tree	0.70 (+/- 0.01)	0.70 (+/- 0.02)	0.70 (+/- 0.01)	0.70 (+/- 0.02)
Random Forest	0.71 (+/- 0.02)	0.71 (+/- 0.02)	0.71 (+/- 0.02)	0.71 (+/- 0.03)
Multilayer Perceptron	0.70 (+/- 0.00)	0.70 (+/- 0.02)	0.71 (+/- 0.02)	0.71 (+/- 0.02)
Naive Bayes	0.67 (+/- 0.01)	0.67 (+/- 0.01)	0.66 (+/- 0.01)	0.67 (+/- 0.02)

Linear SVM	0.69 (+/- 0.02)	0.69 (+/- 0.02)	0.68 (+/- 0.02)	0.68 (+/- 0.03)

Table 6: Top 10 tree root node frequency in the study data

Root Node	Frequency in data	% Frequency
S	48000	55.9%
INTJ	28879	33.6%
S-UNF	3764	4.4%
SQ	1891	2.2%
FRAG	1586	1.8%
NP	698	0.8%
SBAR-PRP	549	0.6%
X	236	0.3%
S-IMP	175	0.2%
PP	119	0.1%

Table 7: 10-fold cross-validation results mean results for top 10 dialogue acts, with features added 1-by-1

	Decision Tree	Random Forest	MLP	Naive Bayes	Linear SVM
BoW	0.63 (+/- 0.02)	0.64 (+/- 0.02)	0.63 (+/- 0.02)	0.63 (+/- 0.01)	0.64 (+/- 0.02)
BoW + Root Node	0.70 (+/- 0.02)	0.71 (+/- 0.03)	0.70 (+/- 0.02)	0.65 (+/- 0.01)	0.64 (+/- 0.02)
BoW + Root Node + Length	0.70 (+/- 0.02)	0.71 (+/- 0.03)	0.71 (+/- 0.02)	0.66 (+/- 0.02)	0.68 (+/- 0.02)
All Features	0.70 (+/- 0.02)	0.71 (+/- 0.03)	0.71 (+/- 0.02)	0.67 (+/- 0.02)	0.68 (+/- 0.03)