A Data-driven Model for Timing Feedback in a Map Task Dialogue System

Raveesh Meena **Gabriel Skantze** **Joakim Gustafson**
KTH Speech, Music and Hearing
Stockholm, Sweden
raveesh@csc.kth.se, gabriel@speech.kth.se, jocke@speech.kth.se

Abstract

We present a data-driven model for detecting suitable response locations in the user's speech. The model has been trained on human–machine dialogue data and implemented and tested in a spoken dialogue system that can perform the Map Task with users. To our knowledge, this is the first example of a dialogue system that uses automatically extracted syntactic, prosodic and contextual features for online detection of response locations. A subjective evaluation of the dialogue system suggests that interactions with a system using our trained model were perceived significantly better than those with a system using a model that made decisions at random.

1 Introduction

Traditionally, dialogue systems have rested on a very simple model for turn-taking, where the system uses a fixed silence threshold to detect the end of the user's utterance, after which the system responds. However, this model does not capture human-human dialogue very accurately; sometimes a speaker just hesitates and no turn-change is intended, sometimes the turn changes after barely any silence (Sacks et al., 1974). Therefore, such models can result in systems that interrupt the user or are perceived as unresponsive. Related to the problem of turn-taking is that of *backchannels* (Yngve, 1970). Backchannel feedback – short acknowledgements such as *uh-huh* or *mm-hm* – are used by human interlocutors to signal continued attention to the speaker, without claiming the floor. If a dialogue system should be able to manage smooth turn-taking and back-channelling, it must be able to first identify suitable locations in the user's speech to do so.

Duncan (1972) found that human interlocutors continuously monitor several cues, such as content, syntax, intonation, paralanguage, and body motion, in parallel to manage turn-taking. Similar observations have been made in various other studies investigating the turn-taking and back-channelling phenomena in human conversations. Ward (1996) has suggested that a low pitch region is a good cue that backchannel feedback is appropriate. On the other hand, Koiso et al. (1998) have argued that both syntactic and prosodic features make significant contributions in identifying turn-taking and back-channelling relevant places. Cathcart et al. (2003) have shown that syntax in combination with pause duration is a strong predictor for backchannel *continuers*. Gravano & Hirschberg (2009) observed that the likelihood of occurrence of a backchannel increases with the number of syntactic and prosodic cues conjointly displayed by the speaker.

However, there is a general lack of studies on how such models could be used online in dialogue systems and to what extent that would improve the interaction. There are two main problems in doing so. First, the data used in the studies mentioned above are from human–human dialogue and it is not obvious to what extent the models derived from such data transfers to human–machine dialogue. Second, many of the features used were manually extracted. This is especially true for the transcription of utterances, but several studies also rely on manually annotated prosodic features.

In this paper, we present a data-driven model of what we call *Response Location Detection* (RLD), which is fully online. Thus, it only relies

Proceedings of the SIGDIAL 2013 Conference, pages 375–383,
Metz, France, 22-24 August 2013.

on automatically extractable features—covering syntax, prosody and context. The model has been trained on human–machine dialogue data and has been implemented in a dialogue system that is in turn evaluated with users. The setting is that of a Map Task, where the user describes the route and the system may respond with for example acknowledgements and clarification requests.

2 Background

Two influential theories that have examined the turn-taking mechanism in human conversations are the signal-based mechanism of Duncan (1972) and the rule-based mechanism proposed by Sacks (1974). According to Duncan, "the turn-taking mechanism is mediated through signals composed of clear-cut behavioural cues, considered to be perceived as discrete". Duncan identified six discrete behavioural cues that a speaker may use to signal the intent to yield the turn. These behavioural cues are: (i) any deviation from the sustained intermediate pitch level; (ii) drawl on the final syllable of a terminal clause; (iii) termination of any hand gesticulation or the relaxation of tensed hand position—during a turn; (iv) a stereotyped expression with *trailing off* effect; (v) a drop in pitch and/or loudness; and (vi) completion of a grammatical clause. According to the rule-based mechanism of Sacks (1974) turn-taking is regulated by applying rules (e.g. "one party at a time") at Transition-Relevance Places (TRPs)—possible completion points of basic units of turns, in order to minimize gaps and overlaps. The basic units of turns (or turn-constructional units) include sentential, clausal, phrasal, and lexical constructions.

Duncan (1972) also suggested that speakers may display behavioural cues either singly or together, and when displayed together they may occur either simultaneously or in tight sequence. In his analysis, he found that the likelihood that a listener attempts to take the turn is higher when the cues are conjointly displayed across the various modalities.

While these theories have offered a function-based account of turn-taking, another line of research has delved into corpora-based techniques to build models for detecting turn-transition and feedback relevant places in speaker utterances.

Ward (1996) suggested that a 110 millisecond (ms) region of low pitch is a fairly good predictor for back-channel feedback in casual conversational interactions. He also argued that more obvious factors, such as utterance end, rising intonation, and specific lexical items, account for less than they seem to. He contended that prosody alone is sometimes enough to tell you what to say and when to say.

In their analysis of turn-taking and backchannels based on prosodic and syntactic features, in Japanese Map Task dialogs, Koiso et al. (1998) observed that some part-of-speech (POS) features are strong syntactic cues for turn-change, and some others are strongly associated with no turn-change. Using manually extracted prosodic features for their analysis, they observed that falling and rising F0 patterns are related to changes of turn, and flat, flat-fall and rise-fall patterns are indications of the speaker continuing to speak. Extending their analysis to backchannels, they asserted that syntactic features, such as filled pauses, alone might be sufficient to discriminate when back-channelling is inappropriate, whereas presence of backchannels is always preceded by certain prosodic patterns.

Cathcart et al. (2003) presented a shallow model for predicting the location of backchannel *continuers* in the HCRC Map Task Corpus (Anderson et al., 1991). They explored features such as POS, word count in the preceding speaker turn, and silence pause duration in their models. A model based on silence pause only inserted a backchannel in every speaker pause longer than 900 ms and performed better than a word model that predicted a backchannel every seventh word. A tri-gram POS model predicted that nouns and pronouns before a pause are the two most important cues for predicting backchannel continuers. The combination of the tri-gram POS model and pause duration model offered a five-fold improvement over the others.

Gravano & Hirschberg (2009) investigated whether backchannel-inviting cues differ from turn-yielding cues. They examined a number of acoustic features and lexical cues in the speaker utterances preceding smooth turn-changes, backchannels, and holds. They have identified six measureable events that are strong predictors of a backchannel at the end of an *inter-pausal unit*: (i) a final rising intonation; (ii) a higher intensity level; (iii) a higher pitch level; (iv) a final POS bi-gram equal to 'DT NN', 'JJ NN', or 'NN NN'; (v) lower values of noise-to-harmonic ratios; and (vi) a longer IPU duration. They also observed that the likelihood of a backchannel increases in quadratic fashion with the number of cues conjointly displayed by the speaker.

When it comes to using these features for making turn-taking decisions in dialogue sys-

tems, there is however, very little related work. One notable exception is Raux & Eskenazi (2008) who presented an algorithm for dynamically setting *endpointing* silence thresholds based on features from discourse, semantics, prosody, timing, and speaker characteristics. The model was also applied and evaluated in the Let's Go dialogue system for bus timetable information. However, that model only predicted the endpointing threshold based on the previous interaction up to the last system utterance, it did not base the decision on the current user utterance to which the system response is to be made.

In this paper, we train a model for online Response Location Detection that makes a decision whether to respond at every point where a very short silence (200 ms) is detected. The model is trained on human–machine dialogue data taken from a first set of interactions with a system that used a very naïve policy for Response Location Detection. The trained model is then applied to the same system, which has allowed us to evaluate the model online in interaction with users.

3 A Map Task dialogue system

In a previous study, we presented a fully automated spoken dialogue system that can perform the Map Task with a user (Skantze, 2012). Map Task is a common experimental paradigm for studying human-human dialogue, where one subject (the information *giver*) is given the task of describing a route on a map to another subject (the information *follower*). In our case, the user acts as the giver and the system as the follower. The choice of Map Task is motivated partly because the system may allow the user to keep the initiative during the whole dialogue, and thus only produce responses that are not intended to take the initiative, most often some kind of feedback. Thus, the system might be described as an *attentive listener*.

Implementing a Map Task dialogue system with full speech understanding would indeed be a challenging task, given the state-of-the-art in automatic recognition of conversational speech. In order to make the task feasible, we have implemented a trick: the user is presented with a map on a screen (see Figure 1) and instructed to move the mouse cursor along the route as it is being described. The user is told that this is for logging purposes, but the real reason for this is that the system tracks the mouse position and thus knows what the user is currently talking about. It is thereby possible to produce a coherent system behaviour without any speech recognition at all, only basic speech detection. This often results in a very realistic interaction, as compared to what users are typically used to when interacting with dialogue systems—in our experiments, several users first thought that there was a hidden operator behind it[1].

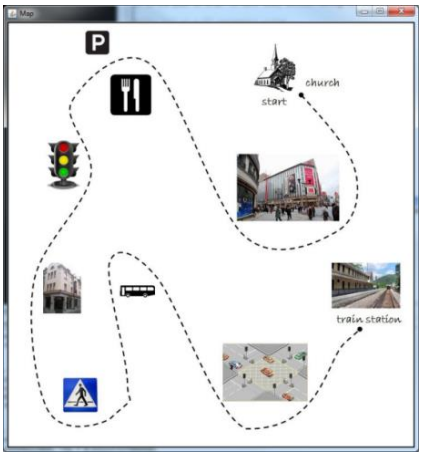

Figure 1: The user interface, showing the map.

The basic components of the system can be seen in Figure 2. Dashed lines indicate components that were not part of the first iteration of the system (used for data collection), but which have been used in the model presented and evaluated here. The system uses a simple energy-based speech detector to chunk the user's speech into inter-pausal units (IPUs), that is, periods of speech that contain no sequence of silence longer than 200 ms. Such a short threshold allows the system to give backchannels (seemingly) while the user is speaking or take the turn with barely any gap. Similar to Gravano & Hirschberg (2009) and Koiso et al. (1998), we define the end of an IPU as a candidate for the Response Location Detection model to identify as a Response Location (RL). We use the term *turn* to refer to a sequence of IPUs which do not have any responses between them.

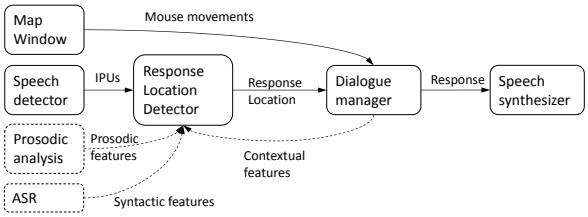

Figure 2: The basic components of the system.

[1] An example video can be seen at http://www.youtube.com/watch?v=MzL-B9pVbOE.

Each time the RLD model detected a RL, the dialogue manager produced a Response, depending on the current state of the dialogue and the position of the mouse cursor. Table 1 shows the different types of responses the system could produce. The dialogue manager always started with an Introduction and ended with an Ending, once the mouse cursor had reached the destination. Between these, it selected from the other responses, partly randomly, but also depending on the length of the last user turn and the current mouse location. Longer turns often led to Restart or Repetition Requests, thus discouraging longer sequences of speech that did not invite the system to respond. If the system detected that the mouse had been at the same place over a longer time, it pushed the task forward by making a Guess response. We also wanted to explore other kinds of feedback than just backchannels, and therefore added short Reprise Fragments and Clarification Requests (see for example Skantze (2007) for a discussion on these).

Table 1: Different responses from the system

Introduction	"Could you help me to find my way to the train station?"
Backchannel	"Yeah", "Mhm", "Okay", "Uhu"
Reprise Fragment	"A station, yeah"
Clarification Request	"A station?"
Restart	"Eh, I think I lost you at the hotel, how should I continue from there?"
Repetition Request	"Sorry, could you take that again?"
Guess	"Should I continue above the church?"
Ending	"Okay, thanks a lot."

A naïve version of the system was used to collect data. Since we initially did not have any sophisticated model of RLD, it was simply set to wait for a random period between 0 and 800 ms after an IPU ended. If no new IPUs were initiated during this period, a RL was detected, resulting in random response delays between 200 and 1000 ms. Ten subjects participated in the data collection. Each subject did 5 consecutive tasks on 5 different maps, resulting in a total of 50 dialogues.

Each IPU in the corpus was manually annotated into three categories: Hold (a response would be inappropriate), Respond (a response is expected) and Optional (a response would not be inappropriate, but it is perfectly fine not to respond). Two human-annotators labelled the corpus separately. For all the three categories the kappa score was 0.68, which is substantial agreement (Landis & Koch, 1977). Since only 2.1% of all the IPUs in the corpus were identified for category Optional, we excluded them from the corpus and focused on the Respond and Hold categories only. The data-set contains 2272 IPUs in total; the majority of which belong to the class Respond (50.79%), which we take as our majority class baseline. Since the two annotators agreed on 87.20% of the cases, this can be regarded as an approximate upper limit for the performance expected from a model trained on this data.

In (Skantze, 2012), we used this collected data to build an offline model of RLD that was trained on prosodic and contextual features. In this paper, we extend this work in three ways. First, we bring in Automatic Speech Recognition (ASR) for adding syntactic features to the model. Second, the model is implemented as a module in the dialogue system so that it can extract the prosodic features online. Third, we evaluate the performance of our RLD model against a baseline system that makes a random choice, in a dialogue system interacting with users.

In contrast to some related work (e.g. Koiso et al., 1998), we do not discriminate between locations for backchannels and turn-changes. Instead, we propose a general model for response location detection. The reason for this is that the system mostly plays the role of an attentive listener that produces utterances that are not intended to take the initiative or claim the floor, but only to provide different types of feedback (cf. Table 1). Thus, suitable response locations will be where the user invites the system to give feedback, regardless of whether the feedback is simply an acknowledgement that encourages the system to continue, or a clarification request. Moreover, it is not clear whether the acknowledgements the system produces in this domain should really be classified as backchannels, since they do not only signal continued attention, but also that some action has been performed (cf. Clark, 1996). Indeed, none of the annotators felt the need to mark relevant response locations within IPUs.

4 A data-driven model for response location detection

The human–machine Map Task corpus described in the previous section was used for training a new model of RLD. We describe below how we extracted prosodic, syntactic and contextual features from the IPUs. We test the contribution of these feature categories—individually as well as

in combination, in classifying a given IPU as either Respond or Hold type. For this we explore the Naïve Bayes (NB) and Support Vector Machine (SVM) algorithms in the WEKA toolkit (Hall et al., 2009). All results presented here are based on 10-fold cross-validation.

4.1 Prosodic features

Pitch and intensity (sampled at 10 ms) for each IPU were extracted using ESPS in Wavesurfer/Snack (Sjölander & Beskow, 2000). The values were transformed to log scale and z-normalized for each user. The final 200 ms voiced region was then identified for each IPU. For this region, the **mean pitch**, **slope of the pitch** (using linear regression)—in combination with the **correlation coefficient *r*** for the regression line, were used as features. In addition to these, we also used the **duration** of the voiced region as a feature. The last 500 ms of each IPU were used to obtain the **mean intensity** (also z-normalised). Table 2 illustrates the power of prosodic features, individually as well as collectively (last row), in classifying an IPU as either Respond or Hold type. Except for mean intensity all other features individually provide an improvement over the baseline. The best accuracy, 64.5%, was obtained by the SVM algorithm using all the prosodic features. This should be compared against the baseline of 50.79%.

Table 2: Percentage accuracy of prosodic features in detecting response locations

	Algorithm	
Feature(s)	**NB**	**SVM**
Mean pitch	60.3	62.7
Pitch slope	59.0	57.8
Duration	58.1	55.6
Mean intensity	50.3	52.2
Prosody (all combined)	63.3	**64.5**

4.2 Syntactic features

As lexico-syntactic features, we use the **word form** and **part-of-speech tag** of the last two words in an IPU. All the IPUs in the Map Task corpus were manually transcribed. To obtain the part-of-speech tag we used the LBJ toolkit (Rizzolo & Roth, 2010). Column three in Table 3 illustrates the discriminatory power of syntactic features—extracted from the manual transcription of the IPUs. Using the last two words and their POS tags, the Naïve Bayes learner achieves the best accuracy of 83.6% (cf. row 7). While POS tag is a generic feature that would enable the model to generalize, using word form as a feature has the advantage that some words, such as *yeah*, are strong cues for predicting the Respond class, whereas pause fillers, such as *ehm*, are strong predictors of the Hold class.

Table 3: Percentage accuracy of syntactic features in detecting response locations

		Manual transcriptions		ASR results	
#	**Feature(s)**	**NB**	**SVM**	**NB**	**SVM**
1	Last word (Lw)	82.5	83.9	80.8	80.9
2	Last word part-of-speech (Lw-POS)	79.4	79.5	74.5	74.6
3	Second last word (2ndLw)	68.1	67.7	67.1	67.0
4	Second last word Part-of-speech (2ndLw-POS)	66.9	66.5	65.8	66.1
5	Lw + 2ndLw	82.3	81.5	80.8	80.6
6	Lw-POS + 2ndLw-POS	80.3	80.5	75.4	74.87
7	Lw + 2ndLw + Lw-POS + 2ndLw-POS	**83.6**	81.7	79.7	79.7
8	Last word dictionary (Lw-Dict)	83.4	83.4	78.0	78.0
9	Lw-Dict + 2ndLw-Dict	81.2	82.6	76.1	77.7
10	Lw + 2ndLw + Lw-Conf + 2ndLw-Conf	82.3	81.5	**81.1**	80.5

An RLD model for online predictions requires that the syntactic features are extracted from the output of a speech recogniser. Since speech recognition is prone to errors, an RLD model trained on manual transcriptions alone would not be robust when making predictions in noisy data. Therefore we train our RLD model on actual speech recognised results. To achieve this, we did an 80-20 split of the Map Task corpus into training and test sets respectively. The transcriptions of IPUs in the training set were used to train the language model of the Nuance 9 ASR system. The audio recordings of the IPUs in the test set were then recognised by the trained ASR system. After performing five iterations of splitting, training and testing, we had obtained the speech recognised results for all the IPUs in the Map Task corpus. The mean word error rate for the five iterations was 17.22% (*SD* = 3.8%).

Column four in Table 3 illustrates the corresponding performances of the RLD model trained on syntactic features extracted from the best speech recognized hypotheses for the IPUs. With the introduction of a word error rate of 17.22%, the performances of all the models us-

ing only POS tag feature decline. The performances are bound to decline further with increase in ASR errors. This is because the POS tagger itself uses the left context to make POS tag predictions. With the introduction of errors in the left context, the tagger's accuracy is affected, which in turn affects the accuracy of the RLD models. However, this decline is not significant for models that use word form as a feature. This suggests that using context independent lexico-syntactic features would still offer better performance for an online model of RLD. We therefore also created a word class **dictionary**, which generalises the words into domain-specific classes in a simple way (much like a class-based n-gram model)**.** Row 9 in Table 3 illustrates that using a dictionary instead of POS tag (cf. row 6) improves the performance of the online model. We have also explored the use of word-level **confidence scores (Conf)** from the ASR as another feature that could be used to reinforce a learning algorithm's confidence in trusting the recognised words (cf. row 10 in Table 3).

The best accuracy, 81.1%, for the *online* model of RLD is achieved by the Naïve Bayes algorithm using the features word form and confidence score, for last two words in an IPU.

4.3 Contextual features

We have explored three discourse context features: **turn** and **IPU length** (in words and seconds) and **last system dialogue act**. Dialogue act history information have been shown to be vital for predicting a listener response when the speaker has just responded to the listener's clarification request (Koiso et al. (1998); Cathcart et al. 2003; Gravano & Hirschberg (2009); Skantze, 2012). To verify if this rule holds in our corpus, we extracted turn length and dialogue act labels for the IPUs, and trained a J48 decision tree learner. The decision tree achieved an accuracy of 65.7%. One of the rules learned by the decision tree is: *if the last system dialogue act is Clarification or Guess (cf. Table 1), and the turn word count is less than equal to 1, then Respond*. In other words, if the system had previously sought a clarification, and the user has responded with a yes/no utterance, then a system response is expected. A more general rule in the decision tree suggests that: *if the last system dialogue act was a Restart or Repetition Request, and if the turn word count is more than 4 then Respond otherwise Hold.* In other words, the system should wait until it gets some *amount* of information from the user.

Table 4 illustrates the power of these contextual features in discriminating IPUs, using the NB and the SVM algorithms. All the features individually provide improvement over the baseline of 50.79%. The best accuracy, 64.8%, is achieved by the SVM learner using the features *last system dialogue act* and *turn word count*.

Table 4: Percentage accuracy of contextual features in detecting response locations

	Manual transcriptions		**ASR results**	
Features	**NB**	**SVM**	**NB**	**SVM**
Last system dialogue act	54.1	54.1	54.1	54.1
Turn word count	61.8	61.9	61.5	62.9
Turn length in seconds	58.4	58.8	58.4	58.8
IPU word count	58.4	58.2	58.1	59.3
IPU length in seconds	57.3	61.2	57.3	61.2
Last system dialogue act + Turn word count	59.9	**64.5**	60.4	**64.8**

4.4 Combined model

Table 5 illustrates the performances of the RLD model using various feature category combinations. It could be argued that the discriminatory power of prosodic and contextual feature categories is comparable. A model combining prosodic and contextual features offers an improvement over their individual performances. Using the three feature categories in combination, the Naïve Bayes learner provided the best accuracy: 84.6% (on transcriptions) and 82.0% (on ASR output). These figures are significantly better than the majority class baseline of 50.79% and approach the expected upper limit of 87.20% on the performance.

Table 5: Percentage accuracy of combined models

	Manual transcriptions		**ASR results**	
Feature categories	**NB**	**SVM**	**NB**	**SVM**
Prosody	63.3	64.5	63.3	64.5
Context	59.9	64.5	60.4	64.8
Syntax	82.3	81.5	81.1	80.5
Prosody + Context	67.7	70.2	67.5	69.1
Prosody + Context + Syntax	**84.6**	77.2	**82.0**	77.1

Table 6 illustrates that the Naïve Bayes model for Response Location Detection trained on combined syntactic, prosodic and contextual features, offers better precision (fraction of correct decisions in all model decisions) and recall (fraction of all relevant decisions correctly made) in comparison to the SVM model.

Table 6: Precision and Recall scores of the NB and the SVM learners trained on combined prosodic, contextual and syntactic features.

Prediction class	Precision (in %)		Recall (in %)	
	NB	SVM	NB	SVM
Respond	81.0	73.0	87.0	84.0
Hold	85.0	81.0	78.0	68.0

5 User evaluation

In order to evaluate the usefulness of the combined model, we have performed a user evaluation where we test the trained model in the Map Task dialogue system that was used to collect the corpus (cf. section 3). A version of the dialogue system was created that uses a Random model, which makes a random choice between Respond and Hold. The Random model thus approximates our majority class baseline (50.79% for Respond). Another version of the system used the Trained model – our data-driven model – to make the decision. For both models, if the decision was a Hold, the system waited 1.5 seconds and then responded anyway if no more speech was detected from the user.

We hypothesize that since the Random model makes random choices, it is likely to produce false-positive responses (resulting in overlap in interaction) as well as false-negative responses (resulting in gap/delayed response) in equal proportion. The Trained model on the other hand would produce fewer overlaps and gaps.

In order to evaluate the models, 8 subjects (2 female, 6 male) were asked to perform the Map Task with the two systems. Each subject performed five dialogues (which included 1 trial and 2 tests) with each version of the system. This resulted in 16 test dialogues each for the two systems. The trial session was used to allow the users to familiarize themselves with the dialogue system. Also, the audio recording of the users' speech from this session was used to normalize the user pitch and intensity for the online prosodic extraction. The order in which the systems and maps were presented to the subjects was varied over the subjects to avoid any ordering effect in the analysis.

The 32 dialogues from the user evaluation were, on average, 1.7 min long (*SD* = 0.5 min). The duration of the interactions with the Random and the Trained model were not significantly different. A total of 557 IPUs were classified by the Random model whereas the Trained model classified 544 IPUs. While the Trained model classified 57.7% of the IPUs as Respond type the Random model classified only 48.29% of the total IPUs as Respond type, suggesting that the Random model was somewhat quieter.

It turned out that it was very hard for the subjects to perform the Map Task and at the same time make a valid subjective comparison between the two versions of the system, as we had initially intended. Therefore, we instead conducted another subjective evaluation to compare the two systems. We asked subjects to listen to the interactions and press a key whenever a system response was either lacking or inappropriate. The subjects were asked not to consider *how* the system actually responded, only evaluate the timing of the response.

Eight users participated in this subjective judgment task. Although five of these were from the same set of users who had performed the Map Task, none of them got to judge their own interactions. The judges listened to the Map Task interactions in the same order as the users had interacted, including the trial session. Whereas it had been hard for the subjects who participated in the dialogues to characterize the two versions of the system, almost all of the judges could clearly tell the two versions apart. They stated that the Trained system provided for a smooth flow of dialogue. The timing of the IPUs was aligned with the timing of the judges' key-presses in order to measure the numbers of IPUs that had been given inappropriate response decisions. The results show that for the Random model, 26.75% of the RLD decisions were perceived as inappropriate, whereas only 11.39% of the RLD decisions for the Trained model were perceived inappropriate. A two-tailed two-sample t-test for difference in mean of the fraction of inappropriate instances (key-press count divided by IPU count) for Random and Trained model show a clear significant difference ($t = 4.66$, $dF = 30$, $p < 0.001$).

We have not yet analysed whether judges penalized false-positives or false-negatives to a larger extent, this is left to future work. However, some judges informed us that they did not penalize delayed response (false-negative), as the system eventually responded after a delay. In the context of a system trying to follow a route description, such delays could sometimes be expected and wouldn't be unnatural. For other types of interactions (such as story-telling), such delays may on the other hand be perceived as unresponsive. Thus, the balance between false-positives and false-negatives might need to be tuned depending on the topic of the conversation.

6 Conclusion

We have presented a data-driven model for detecting response locations in the user's speech. The model has been trained on human–machine dialogue data and has been integrated and tested in a spoken dialogue system that can perform the Map Task with users. To our knowledge, this is the first example of a dialogue system that uses automatically extracted syntactic, prosodic and contextual features for making *online* detection of response locations. The models presented in earlier works have used only prosody (Ward, 1996), or combinations of syntax and prosody (Koiso et al., 1998), syntax and context (Cathcart et al., 2003), prosody and context (Skantze, 2012), or prosody, context and semantics (Raux & Eskenazi (2008). Furthermore, we have evaluated the usefulness of our model by performing a user evaluation of a dialogue system interacting with users. None of the earlier models have been tested in user evaluations.

The significant improvement of the model gained by adding lexico-syntactic features such as word form and part-of-speech tag corroborates with earlier observations about the contribution of syntax in predicting response location (Koiso et al., 1998; Cathcart et al., 2003; Gravano & Hirschberg, 2009). While POS tag alone is a strong generic feature for making predictions in offline models its contribution to decision making in online models is reduced due to speech recognition errors. This is because the POS tagger itself uses the left context to make predictions, and is not typically trained to handle noisy input. We have shown that using only the word form or a dictionary offers a better performance despite speech recognition errors. However, this of course results in a more domain-dependent model.

Koiso et al., (1998), have shown that prosodic features contribute almost as strongly to response location prediction as the syntactic features. We do not find such results with our model. This difference could be partly attributed to inter-speaker variation in the human–machine Map Task corpus used for training the models. All the users who participated in the corpus collection were non-native speakers of English. Also, our algorithm for extracting prosodic features is not as powerful as the manual extraction scheme used in (Koiso et al., 1998). Although prosodic and contextual features do not seem to improve the performance very much when syntactic features are available, they are clearly useful when no ASR is available (70.2% as compared to the baseline of 50.79%).

The subjective evaluation indicates that the interactions with a system using our trained model were perceived as smoother (more accurate responses) as compared to a system using a model that makes a random choice between Respond and Hold.

7 Future work

Coordination problems in turn-transition and responsiveness have been identified as important short-comings of turn-taking models in current dialogue systems (Ward et al., 2005). In continuation of the current evaluation exercise, we would next evaluate our Trained model—on an objective scale, in terms of its responsiveness and smoothness in turn-taking and back-channels. An objective measure is the proportion of judge key-presses coinciding with false-positive and false-negative model decisions. We argue that in comparison to the Random model our Trained model produces (i) fewer instances of false-negatives (gap/delayed response) and therefore has a faster response time, and (ii) fewer instances of false-positives (overlap) and thus provides for smooth turn-transitions.

We have so far explored syntactic, prosodic and contextual features for predicting response location. An immediate extension to our model would be to bring semantic features in the model. In Meena et al. (2012) we have presented a data-driven method for semantic interpretation of verbal route descriptions into *conceptual route graphs*—a semantic representation that captures the semantics of the way human structure information in route descriptions. Another possible extension is to situate the interaction in a face-to-face Map Task between a human and a robot and add features from other modalities such as gaze.

In a future version of the system, we do not only want to determine *when* to give responses but also *what* to respond. In order to do this, the system will need to extract the semantic concepts of the route directions (as described above) and utilize the confidence scores from the spoken language understanding component in order to select between different forms of clarification requests and acknowledgements.

Acknowledgments

This work is supported by the Swedish research council (VR) project *Incremental processing in multimodal conversational systems* (2011-6237).

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4), 351-366.

Cathcart, N., Carletta, J., & Klein, E. (2003). A shallow model of backchannel continuers in spoken dialogue. In *10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest.

Clark, H. H. (1996). *Using language.* Cambridge, UK: Cambridge University Press.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology, 23*(2), 283-292.

Gravano, A., & Hirschberg, J. (2009). Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of Interspeech 2009* (pp. 1019-1022). Brighton, U.K.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1).

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech, 41*, 295-321.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Meena, R., Skantze, G., & Gustafson, J. (2012). A Data-driven Approach to Understanding Spoken Route Directions in Human-Robot Dialogue. In *Proceedings of Interspeech*. Portland, OR, US.

Raux, A., & Eskenazi, M. (2008). Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of SIGdial 2008*. Columbus, OH, USA.

Rizzolo, N., & Roth, D. (2010). Learning Based Java for Rapid Development of NLP Systems. *Language Resources and Evaluation*.

Sacks, H., Schegloff, E., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Sjölander, K., & Beskow, J. (2000). WaveSurfer - an open source speech tool. In Yuan, B., Huang, T., & Tang, X. (Eds.), *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing* (pp. 464-467). Beijing.

Skantze, G. (2007). *Error Handling in Spoken Dialogue Systems - Managing Uncertainty, Grounding and Miscommunication*. Doctoral dissertation, KTH, Department of Speech, Music and Hearing.

Skantze, G. (2012). A Testbed for Examining the Timing of Feedback using a Map Task. In *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*. Portland, OR.

Ward, N., Rivera, A., Ward, K., & Novick, D. (2005). Root causes of lost time and user stress in a simple dialog system. In *Proceedings of Interspeech 2005*. Lisbon, Portugal.

Ward, N. (1996). Using prosodic clues to decide when to produce backchannel utterances. In *Proceedings of the fourth International Conference on Spoken Language Processing* (pp. 1728-1731). Philadelphia, USA.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (pp. 567-578). Chicago.