Improving Unsegmented Statistical Dialogue Act Labelling *

Vicent Tamarit, Carlos-D. Martínez-Hinarejos and José-M. Benedí Ruíz

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
Camino de Vera s/n, 46022, Valencia, Spain
{vtamarit,cmartine,jbenedi}@iti.upv.es

Abstract

An important part of a dialogue system is the correct labelling of turns with dialogue-related meaning. This meaning is usually represented by dialogue acts, which give the system semantic information about user intentions. Each dialogue act gives the semantic of a segment of a turn, which can be formed by several segments. Probabilistic models that perform dialogue act labelling can be used on segmented or unsegmented turns. The last option is the more realistic one, but provides poorer results. An hypothesis on the number of segments can be provided in this case to improve the results. We propose some methods to estimate the probability of the number of segments based on the transcription of the turn. The new labelling model includes the estimation of the probability of the number of segments in the turn. The results show that this inclusion significantly improves the labelling accuracy.

Keywords

dialogue systems, dialogue act, statistical labelling

1 Introduction

A dialogue system is usually defined as a computer system that interacts with a human user to achieve a task using dialogue [6]. The computer system must interpret the user input, in order to obtain the meaning and the intention of the user turn. This is needed to give the appropriate answer to the user. The selection of this answer, along with other decisions that the system can take, is guided by the so-called dialogue strategy. This dialogue strategy can be rule-based [8] or data-based [17]. In the rule-based alternative, the dialogue manager selects the set of actions based on a set of production rules, usually implemented by an expert. In the data-based alternative, there are some ways to build the dialogue system. One option is using a dialogue manager whose parameters have been estimated from annotated data using supervised machine learning techniques, but this approach only take into account the strategies seen in the training data. For this reason simulated users [13] and reinforcement learning [15] are also used to obtain a more robust estimation of the dialogue manager parameters.

In either case, the dialogue strategy needs the interpretation of user turns to achieve the aim of the user. This interpretation must only take into account the essential information for the dialogue process, which is usually represented by special labels called Dialogue Acts (DA) [4]. With this approximation, each user turn can be assigned a sequence of DAs, where each DA is associated with non-overlapped sequences of words in the turn. These sequences of words are usually called segments (some authors refers to these sequences as "utterances" [14]). Each segment has an associated DA which defines its dialogue-related meaning (usually the intention, the communicative function, and the important data).

Therefore, the correct assigment of DAs to a user turn is crucial to the correct behaviour of the dialogue system. The DA tagging is a difficult task even for a human being, because similar segments can be labelled with different DAs depending on the context. Moreover, even the identification of the segments in the turn is a difficult task. To speed-up the labelling time, several models have been proposed to perform this assignment. These assignation models can be based on the annotation rules used by human labellers, but in that case it is quite difficult to code all the rules and exceptions and the model is quite rigid. In recent years, probabilistic data-based models have gained importance for this task [10, 14, 11] as they allow an easier implementation and more flexibility than rule-based models (although they require more annotated data).

The probabilistic parameters of these data-based models are estimated from appropriately labelled dialogue corpora. These dialogue corpora provide sets of dialogues that are segmented and annotated with DA labels. In the posterior use of the models, they are applied to non-annotated dialogues to obtain the most likely DA sequence for each turn. Most of the previous work on DA assignation assumed the correct segmentation of the dialogue turns. However, in a real situation, the only data that are available are the dialogue turns, and the segmentation is not available. Fortunately, these models can be easily adapted to the real situation in which segmentation is not available. In this case, the labelling accuracy is lower than that produced over correctly-segmented dialogue turns.

One possible solution for improving the results on unsegmented turns is to obtain a segmentation hypothesis of the turn before applying the DA assigna-

*Work supported by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01 and by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018).

International Conference RANLP 2009 - Borovets, Bulgaria, pages 434–440

tion model, as that proposed in [1]. In that work, the authors propose a segmentation method based on some lexical and prosodic features, which is then used to make the dialogue act classification. The work presented good results but the classification task is limited to 5 classes.

The estimation of the segmentation can be also achieved in a typed dialogue, but, instead of estimating the entire segmentation, another less restricting possibility is to estimate the number of segments of a given turn. Once the estimation is made, the search for the most likely DA sequence is restricted to only having the estimated number of DA. The estimation of the number of segments can be done using the transcriptions of the turns, so it is possible to use it in typed dialogues, where only the text is available, and in spoken dialogues, because it is possible to use the output of an automatic speech recognition system as the input for the DA tagging.

In this paper, we present the formulation of a general probabilistic model of DA assignation that can be applied on the transcription of unsegmented turns. The model evolves from this general formulation to a more restricted formulation where first the probability of the number of segments is estimated, and then the most likely segmentation is obtained. Initial results show that estimating the probability of the number of segments produces significant improvements in the accuracy of the DA assignation. Following this, we present a model to estimate the number of segments given the available dialogue features (words of the turn and its length). The combination of this model with the DA assignation model shows significant improvement in accuracy with respect to the original unsegmented model.

The paper is organised as follows: In Section 2, we present the HMM-based models for labelling the turns. In Section 3, we introduce the estimation of the number of segments and describe the different approaches to the combination of features for that estimation. In Section 4, we present the experiments for testing the models as well as the results. In Section 5 we present our final conclusions and future work.

2 The HMM-based model for DA assignation

Given a word sequence $\mathcal{W}$, the main goal is to obtain the optimum DA sequence $\widehat{\mathcal{U}}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$.

The DA sequence $\mathcal{U}$ of the complete dialogue can be seen as $U_1^{t-1} = U_1 \cdot U_2 \cdots U_{t-1}$, which represents the DA sequence detected until the current turn t. The word sequence of the current turn is expressed as $W = W_1^l = w_1 \cdot w_2 \cdots w_l$, where l is the number of words of W. Therefore, we can reformulate the problem by introducing a new posterior probability $\Pr(U|W_1^l, U_1^{t-1})$, which represents the probability of the DA sequence U that is associated to the current user turn, given the word sequence of the user turn W_1^l and the history of the previous DA sequence U_1^{t-1}. The goal is to find the best sequence of DAs for each turn:

$$\widehat{U} = \underset{U}{\operatorname{argmax}} \Pr(U|W_1^l, U_1^{t-1}) \quad (1)$$

Then, we can introduce two *hidden* variables: the number of segments r; and the segmentation of the turn, which can be described as $s = (s_0, s_1, \ldots, s_r)$. Therefore, U can be expressed as $U = u_1^r$, and W as $W_1^l = W_{s_0+1}^{s_1} W_{s_1+1}^{s_2} \ldots W_{s_{r-1}+1}^{s_r}$.

From Equation (1) we can derive two models. The usual assumption is that the segmentation s and the number of segments r are unknown and have no influence on the DA assignation. In this case, as we are under the argmax framework, we can express the probability of the DA sequence as:

$$\Pr(U|W_1^l, U_1^{t-1}) = \Pr(U|U_1^{t-1}) \Pr(W_1^l|U, U_1^{t-1}) = \sum_{r, s_1^r} \prod_{k=1}^{r} \Pr(u_k|u_1^{k-1}, U_1^{t-1}) \Pr(W_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}) \quad (2)$$

This model is simplified with three basic assumptions: the probability of the word segments depends only on the current DA; the probability of the DA depends only on the n previous DAs; and the summation is replaced by a maximisation. The resulting model is the following:

$$\Pr(U|W_1^l, U_1^{t-1}) = \max_{r, s_1^r} \prod_{k=1}^{r} \Pr(u_k|u_{k-n-1}^{k-1}) \Pr(W_{s_{k-1}+1}^{s_k}|u_k) \quad (3)$$

This model can be used when there is an available segmentation (and consequently we know the correct number of segments r) by simply eliminating the maximisation and fixing the s_k values and r to those provided by the segmentation. If there is no segmentation available, the search for the optimal DA sequence provides a segmentation that allows the maximum probability to be obtained. Consequently, we can obtain a segmentation derived from this method. This model can be considered as the baseline model.

We can develop another model from Equation (1) if we consider a different assumption: the number of segments influences the labelling. In this case, the probability of the sequence U is:

$$\Pr(U|W_1^l, U_1^{t-1}) = \sum_r \Pr(U, r|W_1^l, U_1^{t-1}) = \sum_r \Pr(r|W_1^l, U_1^{t-1}) (\Pr(U|U_1^{t-1}, r) \Pr(W_1^l|U, U_1^{t-1}, r)) = \sum_r \Pr(r|W_1^l, U_1^{t-1}) \prod_{k=1}^{r} \Pr(u_k|u_1^{k-1}, U_1^{t-1}, r) \Pr(W_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}, r) \quad (4)$$

To simplify this expression, we do the same simplifications that we did to obtain Equation (3). Thus, the new labelling model is:

$$\Pr(U|W_1^l, U_1^{t-1}) = \sum_r \Pr(r|W_1^l, U_1^{t-1})$$

$$\max_{s_1^r} \prod_{k=1}^{r} \Pr(u_k|u_{k-n-1}^{k-1}) \Pr(W_{s_{k-1}+1}^{s_k}|u_k) \quad (5)$$

As in the previous model, we can obtain a segmentation from this Equation.

In Equations (3) and (5), $\Pr(u_k|u_{k-n-1}^{k-1})$ can be modelled as an n-gram (of degree n) and $\Pr(W_{s_{k-1}+1}^{s_k}|u_k)$ can be modelled as a HMM. The maximisation (including the segmentation and the DA decoding) can be implemented using the Viterbi algorithm. Note that, in this formula, u_{k-n-1}^{k-1} can take DAs of previous turns.

Therefore, we have derived two labelling models from Equation (1). The model described in Equation (3) does not contain any information about the number of segments of the turn nor any information about the segmentation. The model presented in Equation (5) includes the estimation of the probability of the number of segments.

To estimate the probability $\Pr(r|W_1^l, U_1^{t-1})$, the dependencies of r are substituted by a score S_c that is explained in the next section.

3 Estimation of the number of segments

In Section 2, we introduced an approach to estimate the number of segments of a turn; that is, we defined a score S_c associated with each turn, which is computed from the transcription. To estimate the number of segments, we chose the approximation $\Pr(r|W_1^l, U_1^{t-1}) = \Pr(r|S_c)$, where S_c is calculated in from the sequence of words W_1^l.

The new probability can be calculated by applying the Bayes rule:

$$\Pr(r|S_c) = \frac{p(S_c|r)p(r)}{p(S_c)} \quad (6)$$

The a priori probability $p(r)$ can be easily computed as the number of turns with r segments, N_{Tr}, divided by the total number of turns N_T:

$$p(r) = \frac{N_{Tr}}{N_T} \quad (7)$$

The conditional member $p(S_c|r)$ is estimated by a normal distribution. We calculated one distribution for each r:

$$p(S_c|r) \sim \mathcal{N}(m_r, \sigma_r) \quad (8)$$

The mean m_r and variance σ_r are computed from the scores associated with the turns with r segments.

The last element $P(S_c)$ is estimated by another gaussian distribution that is computed from all the turns:

$$p(S_c) \sim \mathcal{N}(m_{Sc}, \sigma_{Sc}) \quad (9)$$

The mean m_{Sc} and variance σ_{Sc} are computed from all the scores in the training data.

The computation of S_c is made using features that are extracted from the transcription of each turn (it is word-based). We have focused on two features to estimate the number of segments of a turn. One evident feature is the number of words of the turn. More sophisticated features can be inferred from the words (or sequences) that usually appear at the beginning or the end of segments.

First, we made a study of the features that could determine the number of segments and we evaluated the influence of some of them:

- Length of the turn. We evaluated the relation between the number of segments and the number of words in a turn.
- Final words and final n-grams. In the transcription, some words (like the interrogation mark and the period) clearly indicate the end of a segment. Combinations of the last two or three words are also useful.
- Initial words and n-grams. This is the opposite case to the final words.
- Combinations: The above features can be combined to obtain a better estimation of the number of segments.

Second, we defined some calculations for the score S_c based on the above-mentioned features.

- Based on length of the turn

  The score S_c can be calculated as the number of words in the turn:

  $$S_c(W) = l \quad (10)$$

- Boundary words

  We define the score S_c of a turn W as:

  $$S_c(W) = \sum_{i=1}^{l} p_f(w_i) \quad (11)$$

  where $p_f(w_i)$ is the probability of the word i being a final word in a segment. It is estimated by counting in the training corpus the number of times that the word is final divided by the total number of appearances of the word. This value is 0 for the words that never appear at the end of a segment.

  It is also possible to calculate S_c in the same way using the initial words of a segment instead of final ones.

- Boundary n-grams

  Instead of calculating the probability of a final word, we propose the estimation of the probability of the n last words of the segments. In this case, the method of estimation is the same one that we used in the above case: the number of times that the n-gram is at the end of the segment divided by the total number of appearances of the n-gram. We calculated the S_c using that estimation with:

$$S_c(W) = \sum_{i=n}^{l} p_f(W_{i-(n-1)}^{i}) \tag{12}$$

  As we proposed in the final word estimation, the probability of initial n-grams in a segment can be computed just by counting the times an n-gram is initial.

- Composed score

  The features that we used in the estimation of the score can be combined. In this case, the score calculated for a turn is composed of various features, e.g. the score can be seen as the summation of the probability of each word to be final plus the length of the turn (by adding a number a for each word):

$$S_c(W) = \sum_{i=1}^{l} (p_f(w_i) + a) \tag{13}$$

  Another option is to combine the final words with final n-grams, e.g., combining the final bigrams and the final words:

$$S_c(W) = \sum_{i=2}^{l} p_f(W_{i-1)}^{i}) + \sum_{i=1}^{l} p_f(w_i) \tag{14}$$

- Naive-Bayes Score

  In this case, the final probability of the number of segments is calculated by combining the probabilities for each score, i.e., if we consider:

$$\Pr(r|S_{c_1}, S_{c_2}, \cdots S_{c_n})$$

  this probability can be simplified assuming that there are no dependencies between scores (naive-Bayes assumption):

$$\Pr(r|S_{c_1}, S_{c_2}, \cdots S_{c_n}) = \Pr(r|S_{c_1}) \Pr(r|S_{c_2}) \cdots \Pr(r|S_{c_n}) \tag{15}$$

4 Experiments and results

We present three sets of experiments that we performed using the SwitchBoard corpus [7]. The experiments were designed to show the error in the estimation of the number of segments and the accuracy of the labelling provided by the two models described in Section 2 (Equation (3) and Equation (5)).

4.1 SwitchBoard Corpus

The SwitchBoard corpus [7] is a well-known corpus of human-human conversations by telephone. The conversations are not related to a specific task, since the speakers discuss general interest topics, with no clear task to accomplish. This corpus recorded spontaneous speech, with frequent interruptions between the speakers and background noises. The transcription of the corpus takes into account all these facts and it includes special notation for the overlaps, noises and other sound effects present in the recordings.

The corpus is composed of 1,155 different conversations in which 500 different speakers participated. The number of turns in the dialogues is around 115,000, including overlaps. The vocabulary size is approximately 42,000 words.

The corpus was manually divided into segments following the criteria defined by [9], and annotated using a shallow version of the SWBD-DAMSL annotation scheme [5]. Each segment is labelled with one of the 42 different labels present in the SWBD-DAMSL annotation set. These labels represent categories such as statement, backchannel, questions, answers, etc., and different subcategories for each of these categories (e.g., statement opinion/non-opinion, yes-no/open/rethorical-questions, etc.). The manual labelling was performed by 8 different human labellers, with a Kappa value of 0.80, which reflects the difficulty of the segmentation and annotation task. This corpus is generally used in the evaluation of statistical annotation models ([14], [12], [16])

To simplify the labelling task, we-preprocessed the transcriptions of the SwitchBoard corpus to remove certain particularities. The interrupted segments were joined, thereby avoiding interruptions and ignoring overlaps between the speakers. The vocabulary was reduced by using all the words in lowercase and separating the punctuation marks from the words.

To obtain more reliable results, we performed a partition on the corpus to perform experiments with a cross-validation approach. In our case, the 1,155 different dialogues were divided into 11 partitions with 105 dialogues each one.

4.2 Estimation of the number of segments

The first set of experiments were the tests to determine the best way to estimate the number of segments of a turn. Table 1 shows the results of the different estimations of the number of segments.

These tests showed that the final bigrams provided the best estimation of the number of segments. The initial words (or bigrams) did not estimate the number of segments as well as the final ones; even the length of the turn was a better estimator. The final words and n-grams produced better results due to the presence of some words that always indicate the end of a segment (like the interrogation mark and the period). The two kinds of combination (composed and naive-vayes) did not produce any improvement in the estimation.

Estimation	Error
Length	35.8
Final Words	33.4
Final Bigrams	**27.9**
Final Trigrams	37.4
Initial Words	39.1
Initial Bigrams	39.1
Initial Trigrams	39.0
Composed length and final word score	35.6
Naive-Bayes of length and final words	34.8

Table 1: *Results of the estimation of the number of segments. The estimation column indicates the type of the score used in the estimation of r. The error column indicates the percent of the turns where the estimated number of segments is different from the real number of segments.*

4.3 Baseline

In Section 2, we presented two models for labelling. The baseline experiments used the model represented by Equation (3). We estimated the DA Error Rate (DAER) and the Turn Error Rate (TER). The DAER is the average edit distance between the reference DA sequences and the DA sequences assigned by the labelling model. The TER indicates the percent of turns that are incorrectly labelled. Table 2 shows the results using 2-grams and 3-grams for the estimation of the probability $\Pr(u_k|u_{k-n-1}^{k-1})$. It shows a comparison of the error in the labelling between the segmented and the unsegmented version. In the segmented version we knew the correct segmentation, but in the unsegmented version we did not know anything about the segmentation or the number of segments. The segmented version is a hypothetic case, because in a real system we do not know the correct segmentation.

2-gram		
Corpus	DAER/TER	C.I.
Segmented	29.9/38.3	± 0.6
Unsegmented	63.2/59.6	± 0.5

3-gram		
Corpus	DAER/TER	C.I.
Segmented	29.9/38.1	± 0.6
Unsegmented	**62.5/59.0**	± 0.4

Table 2: *DAER and TER baseline results with the model described in Equation (3). The errors are presented for both segmented and unsegmented corpus. The C.I. column indicates the 90% confidence interval of the DAER. The baseline result considered for the next experiments is shown in boldface.*

These results are boundary errors and they are similar to those provided by [12], where the authors proposed a HMM model to dialogue act labelling. The segmented turns gave us the minimum error supplied by the HMM-based model. The unsegmented turns gave us the maximum error, obtained without knowing the segmentation. We consider that the result obtained with the unsegmented version and a 3-gram is a baseline error (62.5% of DAER). This experiment is useful because it allows us to measure the difference between this model and the one with the estimation of the number of segments. We also included a 90% confidence interval for the DAER to ensure statistical significance. This confidence interval is estimated using a bootstrap estimation [3]. We used each partition as a segment for the bootstrapping and the bootstrap sample is composed by 10,000 elements.

4.4 Labelling with the estimation of the number of segments

The third set of experiments shows the labelling of the turns produced by the mathematical model presented in Equation (5), where we introduce an estimation of the probability of the number of segments. Due to the results of the estimation of r, we used the final words, final bigrams, final trigrams and length features as score estimators. We tested the labelling with 2-grams and 3-grams as estimators of the probability $\Pr(u_k|u_{k-n-1}^{k-1})$.

Table 3 shows a comparison of the errors obtained in the experiments. The error with correct r estimation was computed by labelling the unsegmented corpus, knowing the correct number of segments ($\Pr(r|S_c)$ is 1 for the correct r and 0 for the rest). The inclusion of the labelling with the correct r is only for reference, because it represents an hypothetical case. The rest of the lines refer to different estimations of the number of segments.

2-gram		
r estimation	DAER/TER	C.I.
Correct r	47.4/48.1	± 0.6
Length	54.7/54.9	± 0.5
Final Words	54.2/54.2	± 0.5
Final Bigrams	53.6/53.5	± 0.5
Final Trigrams	54.6/54.8	± 0.5

3-gram		
r estimation	DAER/TER	C.I.
Correct r	47.2/48.1	± 0.5
Length	54.6/54.8	± 0.5
Final Words	54.1/54.2	± 0.5
Final Bigrams	**53.5/53.5**	± 0.5
Final Trigrams	54.5/54.8	± 0.4

Table 3: *DAER and TER results of the labelling using the estimation of segments and different n-grams to estimate $\Pr(u_k|u_{k-n-1}^{k-1})$. Each line refers to a different estimation of the number of segments. It includes the labelling error and a 90% confidence interval for the DAER. The inclusion of the labelling with the correct r is only for reference.*

The best result was obtained with the estimation of the number of segments based on final bigrams and the probability of the dialogue act given by a 3-gram. The confidence interval for this experiment and confidence interval of the baseline show that the difference between the results given by the models are statistically significant. Thus, it can be concluded that the model

2-gram			
r estimation	Precision	Recall	F-measure
No estimation	0.44	0.32	0.37
Correct r	0.45	0.45	0.45
F. Bigrams	0.50	0.42	0.46

3-gram			
r estimation	Precision	Recall	F-measure
No estimation	0.44	0.33	0.38
Correct r	0.46	0.46	0.46
F. Bigrams	0.50	0.42	0.45

Table 4: *Precision, recall and F-measure of the labelling. It includes the results of the baseline labelling error (with no estimation), the labelling error with the correct r estimation and the labelling error using bigrams for the estimation of the number of segments.*

with the estimation of the probability of the number of segments produces a significant improvement in the labelling.

The labelling errors show that there is a relation between the estimation error of the number of segments and the labelling; however, the improvement in the estimation of segments is not translated in the same magnitude to the labelling process. This is due to the difficulty of correctly labelling some turns which were not correctly labelled in any of the experiments, even when the correct number of segments is given. As is pointed out in [14], the cause of these errors could be that some DA definitions are arbitrary and may even confuse a human labeller. To corroborate this problem, we calculated the precision, recall and F-measure of the experiments.

The precision is calculated by dividing the number of correct labelled segments by the total number of labels given by the labeller. The recall is calculated by dividing the number of correct labelled segments by the correct number of segments. The F-measure is computed as $F = 2 \cdot (\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall})$.

Table 4 shows the precision, recall and F-measure of some experiments. The precision indicates the accuracy of the labeller, but the position of the labels in the labelling are not important, thus this errors are better than the corresponding DAER. The precision is similar for all the experiments, which means that the errors are produced by the labeller, even with the correct number of segments. The results also show the improvement produced by the inclusion of the probability of the number of segments in the labelling.

5 Conclusions and future work

In this work, we have shown two different models for the labelling of turns in a dialogue. Both of them are text-based methods, so they can be used in typed dialogues or in spoken dialogues with an automatic speech recogniser. One model directly labels the turns without knowing the segmentation or the number of segments in the turn, and the other model assumes the previous estimation of the probability of the number of segments. Some methods to estimate the probability of the number of segments of a turn based on the transcription are also presented.

The results show that the dialogue act labelling task can be improved by including the probability distribution of the number of segments. Even though our best result is not as good as the one obtained using the correct segmentation, it is significantly better than the error of the unsegmented model with no estimation of the number of segments. Furthermore, the estimation of the probability of the number of segments can be easily computed.

Future work is directed to obtaining a better model that estimates the number of segments. However, the estimations based on the transcription of turns does not seem to produce good enough results. In spoken dialogues, a new estimation could be to use features that are extracted directly from the audio signal, as proposed in [1], and include them into our probability model of the estimation of the number of segments. Another possibility is to repeat these experiments using a corpus of a different kind, such as a task-oriented corpus like Dihana [2]. Moreover, the experiments can be made using the output of a speech recogniser or using a modified version of the corpora with no marks such as points or commas.

References

[1] J. Ang, Y. Liu, and E. Shriberg. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processings*, volume 1, pages 1061–1064, Philadelphia, 2005.

[2] J.-M. Benedí, E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. López de Letona, and A. Miguel. Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pages 1636–1639, May 2006.

[3] M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in asr performance evaluation. *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 1:I–409–12 vol.1, May 2004.

[4] H. Bunt. Context and dialogue control. *THINK Quarterly*, 3, 1994.

[5] M. G. Core and J. F.Allen. Coding dialogues with the damsl annotation scheme. In *Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35. American Association for Artificial Intelligence, Nov 1997.

[6] L. Dybkjaer and W. Minker. *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*. Springer, 2008.

[7] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. *Acoustics, Speech, and*

Signal Processing, IEEE International Conference on, 1:517–520, 1992.

[8] A. Gorin, G. Riccardi, and J. Wright. How may i help you? *Speech Communication*, 23:113–127, 1997.

[9] D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard swbd-damsl shallow- discourse-function annotation coders manual - draft 13. Technical Report 97-01, University of Colorado Institute of Cognitive Science, 1997.

[10] L. Levin, K. Ries, A. Thymé-Gobbel, and A. Levie. Tagging of speech acts and dialogue games in Spanish call home. In *Workshop: Towards Standards and Tools for Discourse Tagging*, pages 42–47, 1999.

[11] C. Martinez-Hinarejos, J. Benedí, and R. Granell. Statistical framework for a spanish spoken dialogue corpus. *Speech Communication*, 50:992–1008, 2008.

[12] C. D. Martínez-Hinarejos, R. Granell, and J. M. Benedí. Segmented and unsegmented dialogue-act annotation with statistical dialogue models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sesions*, pages 563–570, Sydney, Australia,, 17th-21th July 2006.

[13] J. Schatzmann, B. Thomson, and S. Young. Statistical user simulation with a hidden agenda. In *Proc. SIGdial Workshop on Discourse and Dialogue*, pages 273–282, 2007.

[14] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modelling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):1–34, 2000.

[15] M. A. Walker. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416, 2000.

[16] N. Webb, M. Hepple, and Y. Wiks. Dialogue act classification using intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, Pittsburgh, USA, 2005.

[17] S. Young. Probabilistic methods in spoken dialogue systems. *Philosophical Trans Royal Society (Series A)*, 358(1769):1389–1402, 2000.