Natural Language Engineering **18** (1): 1–19. © Cambridge University Press 2011
doi:10.1017/S135132491000032X

Estimating the number of segments for improving dialogue act labelling[†]

VICENT TAMARIT, CARLOS-D. MARTÍNEZ-HINAREJOS and JOSÉ-MIGUEL BENEDÍ

Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, Valencia, Spain
e-mail: {vtamarit,cmartine,jbenedi}@iti.upv.es

(*Received 10 March 2010; revised 7 July 2010; accepted 20 October 2010; first published online 14 February 2011*)

Abstract

In dialogue systems it is important to label the dialogue turns with dialogue-related meaning. Each turn is usually divided into segments and these segments are labelled with dialogue acts (DAs). A DA is a representation of the functional role of the segment. Each segment is labelled with one DA, representing its role in the ongoing discourse. The sequence of DAs given a dialogue turn is used by the dialogue manager to understand the turn. Probabilistic models that perform DA labelling can be used on segmented or unsegmented turns. The last option is more likely for a practical dialogue system, but it provides poorer results. In that case, a hypothesis for the number of segments can be provided 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. We tested this new approach with two different dialogue corpora: SwitchBoard and Dihana. The results show that this inclusion significantly improves the labelling accuracy.

1 Introduction

A dialogue system is usually defined as a computer system that interacts with a human user to achieve a task by using dialogue (Dybkjaer and Minker 2008). The computer system must interpret the user input in order to obtain the meaning and the intention of the user turn. This information is needed in order to give the appropriate response 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 (Gorin, Riccardi and Wright 1997) or data-based (Young 2000). In the rule-based alternative, the dialogue manager selects the set of actions based on a set of production rules, which is usually implemented

[†] Work supported by the EC (FEDER/FSE), the Spanish Government (MEC, MICINN, MITyC, MAEC, "Plan E", under grants MIPRCV "Consolider Ingenio 2010" CSD2007-00018, MITTRAL TIN2009-14633-C03-01, erudito.com TSI-020110-2009-439, FPI fellowship BES-2007-16834), and Generalitat Valenciana (grant Prometeo/2009/014 and grant ACOMP/2010/051).

by an expert. In the data-based alternative, there are several ways to build the dialogue system. One option is to use a dialogue manager whose parameters have been estimated from annotated data using supervised machine-learning techniques; however, this approach only takes into account the strategies seen in the training data. For this reason, simulated users (Schatzmann, Thomson and Young 2007) and reinforcement learning (Walker 2000) 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) (Bunt 1994). With this approach, each user turn can be assigned a sequence of DAs, where each DA is associated with nonoverlapped sequences of words in the turn. These sequences of words are usually called segments (some authors refer to these sequences as 'utterances' (Stolcke *et al.* 2000)). Each segment has an associated DA that defines its dialogue-related meaning The DA usually includes the intention, the communicative function, and the relevant information contained in the segment. The relevant information is defined for each dialogue system since depends on the task the system is related to, e.g. in a train information system, the destination city or departure times are considered relevant information.

Therefore, the correct assignation of DAs to a user turn is crucial to the correct behaviour of the dialogue system. DA tagging is a task that is difficult even for a human being because similar segments can be labelled with different DAs depending on the context. Even the identification of the segments in the turn is a difficult task. Thus, to perform the labelling of dialogue turns several automatic models have been proposed. These labelling 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 (Leviv *et al.* 1999; Stolcke *et al.* 2000; Martínez-Hinarejos, Benedí and Granell 2008) since they are easier to implement and offer more flexibility than rule-based models (even though 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 nonannotated turns to obtain the most likely DA sequence. Most of the previous work on DA assignation assumed the correct segmentation of the dialogue turns. However, this assumption is not valid when the DA labelling is used in a real dialogue system, where segmentation is not available and the only available data are the dialogue turns. 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 results on unsegmented turns is to obtain a segmentation hypothesis of the turn before applying the DA assignation model, as proposed in (Ang, Liu and Shriberg 2005). In that work, the authors proposed a segmentation method based on certain lexical and prosodic features, which was

then used to make the DA classification. The features were extracted from dialogue audio and the resulting transcriptions. The authors test the segmentation method with the ICSI-MRDA corpus. The work presented good labelling results but the classification task was limited to five classes.

Instead of estimating the entire segmentation of each turn, another less restrictive possibility is to estimate the number of segments of a given turn. As it was pointed out in (Martínez-Hinarejos 2009), this estimation of the number of segments can guide the search for the most likely DA sequence. The estimation of the number of segments can be done using the transcriptions of the turns. In this case, it is possible to use this estimation in typed dialogues (where only the text is available), as well as in spoken dialogues (because it is possible to use the output of an automatic speech recognition system as the input for DA tagging).

The goal of this work is to label dialogue turns with the correct sequence of DAs. This DA sequence aids the system to understand the input of a user in a dialogue system. In this case, the correct labelling is more important than the segmentation, because the system needs the correct labels, no matters where in the turn they appear. It is similar to what happens in speech recognition, where only the sequence of words is important, since the correct assignation of each word to the correct part of the speech signal is mostly unimportant. Also, estimating the number of segments is faster and possibly more robust than estimating the entire segmentation.

In this paper, we present the formulation of a general probabilistic model of DA assignation that can be applied on the transcripts 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. We compare the labelling produced by this model with the classic labelling model where the number of segments is not estimated. Initial results show that estimating the probability of the number of segments produces significant improvements in the accuracy of the DA assignation. In accordance with this, we present a model to estimate the number of segments given the available dialogue features (the words and the length of the turn). The combination of this model with the DA assignation model shows significant improvement in the DA labelling with respect to the original unsegmented model.

The paper is organised as follows. In Section 2, we present the statistical models for labelling the unsegmented dialogue turns; we develop the classic model with no information about the number of segments and the new model with the estimation. In Section 3, we introduce the estimation of the number of segments. In Section 4, we describe the two corpora that we used to test the models. In Section 5, we present the experiments performed to test the models as well as the results. In Section 6, we present our final conclusions and future work.

2 HMM-based model for DA labelling

Given an entire dialogue, we consider it as a word sequence $\mathscr{W}$. The main goal of the labelling is to obtain the optimum DA sequence $\widehat{\mathscr{U}}$ that maximises the posterior probability $\Pr(\mathscr{U}|\mathscr{W})$, where $\mathscr{U}$ is a sequence of DAs.

This word sequence $\mathscr{W}$ can be divided into T turns $\mathscr{W} = W_1^T = W_1 W_2 \cdots W_T$, where a turn t has a sequence of words W_t. The same decomposition can be applied to the DA sequence $\mathscr{U} = U_1^T = U_1 U_2 \cdots U_T$. Each turn t presents a sequence of DA U_t. Thus, we can express the optimisation problem as

$$\begin{aligned}\widehat{\mathscr{U}} &= \underset{\mathscr{U}}{\operatorname{argmax}} \Pr(\mathscr{U}|\mathscr{W}) = \underset{U_1^T}{\operatorname{argmax}} \Pr\left(U_1^T|W_1^T\right) \\ &= \underset{U_1^T}{\operatorname{argmax}} \prod_{t=1}^{T} \Pr\left(U_t|U_1^{t-1}\right) \Pr\left(W_t|W_1^{t-1}, U_1^T\right) \end{aligned} \tag{1}$$

This approach is useful for the annotation of an entire dialogue; however, in a real dialogue system, the speech recogniser gives one user turn at a time, and we are interested in labelling it with the correct DA sequence. Therefore, we have to develop a labelling model that restricts the optimisation to a given turn t. In this case, the optimisation problem is transformed into

$$\widehat{U}_t = \underset{U_t}{\operatorname{argmax}} \Pr\left(U_1^t|W_1^t\right) \approx \underset{U_t}{\operatorname{argmax}} \Pr\left(U_t|W_1^t, U_1^{t-1}\right) \Pr\left(U_1^{t-1}|W_1^t\right) \tag{2}$$

When we label a user turn in a dialogue system, we assume that the assignation for previous turns is fixed (since this assignation aided the dialogue manager previously, it cannot be changed). Consequently $\Pr(U_1^{t-1}|W_1^t)$ is constant and it can be taken out of the optimisation. Moreover, it is reasonable to suppose that the words of previous turns, W_1^{t-1}, do not have any influence on the DA sequence of the current turn U_t, since their effect is reflected in the sequence of previous DA U_1^{t-1}. To simplify notation, we use $U = U_t$ and $W = W_t$. Thus, the maximisation problem is formulated as:

$$\widehat{U} = \underset{U}{\operatorname{argmax}} \Pr\left(U|W, U_1^{t-1}\right) \tag{3}$$

Since the labelling of the user turn induces a segmentation, 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 can be expressed as $W = w_{s_0+1}^{s_1} w_{s_1+1}^{s_2} \ldots w_{s_{r-1}+1}^{s_r}$, with $s_0 = 0$ and $s_r = |W|$. From (3), we can derive two models: The first model is the classical approach where the segmentation and the number of segments are unknown; the second model is built assuming that the number of segments can be estimated.

2.1 Classic model

The classic model is produced by the assumption that the segmentation s and the number of segments r are unknown and have no influence on the DA assignation. Therefore, since we are under the argmax framework, we can express the probability of the DA sequence as

$$\begin{aligned}\widehat{U} &= \underset{U}{\operatorname{argmax}} \Pr\left(U|W, U_1^{t-1}\right) \approx \underset{U}{\operatorname{argmax}} \Pr\left(U|U_1^{t-1}\right) \Pr\left(W|U, U_1^{t-1}\right) \\ &= \underset{U}{\operatorname{argmax}} \sum_{r, s_1^r} \prod_{k=1}^{r} \Pr\left(u_k|u_1^{k-1}, U_1^{t-1}\right) \Pr\left(w_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}\right) \end{aligned} \tag{4}$$

Notice that we assume that there are no dependencies between the sequence of words of the current turn and the previous DA sequence. Furthermore, we simplify this model with two basic assumptions: the probability of the word segments depends only on the current DA, and the probability of the DA depends only on the n previous DAs. From this equation, we can obtain two models.

The first model is the simplification of (4). In this case, we do not know the segmentation of the turns

$$\begin{aligned}\widehat{U} &= \underset{U}{\operatorname{argmax}} \Pr\left(U|W, U_1^{t-1}\right) \\ &\approx \underset{U}{\operatorname{argmax}} \sum_{r,s_1^r} \prod_{k=1}^{r} \Pr\left(u_k|u_{k-(n-1)}^{k-1}\right) \Pr\left(w_{s_{k-1}+1}^{s_k}|u_k\right)\end{aligned} \tag{5}$$

The second model is the result of having a segmentation available, which implies that we know the correct number of segments of each turn (Stolcke *et al.* 2000). In this case, we can eliminate the summation and fix the s_k values and r to those provided by the segmentation. The model can be rewritten with the correct $\widehat{r}$ and the correct segmentation $\widehat{s}$

$$\begin{aligned}\widehat{U} &= \underset{U}{\operatorname{argmax}} \Pr\left(U|W, U_1^{t-1}\right) \\ &\approx \underset{U}{\operatorname{argmax}} \prod_{k=1}^{\widehat{r}} \Pr\left(u_k|u_{k-(n-1)}^{k-1}\right) \Pr\left(w_{\widehat{s}_{k-1}+1}^{\widehat{s}_k}|u_k\right)\end{aligned} \tag{6}$$

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. Since we want to label unsegmented turns where the segmentation is unknown, we consider the model described by (5) as the baseline model for the DA labelling. We consider the model described by (6) as an optimistic estimation of the labelling model performance.

2.2 Model with the number of segments

From (3), we developed another model by considering a different assumption: the number of segments influences the labelling. In this case, the probability of the sequence U is

$$\begin{aligned}\widehat{U} &= \underset{U}{\operatorname{argmax}} \Pr\left(U|W, U_1^{t-1}\right) = \underset{U}{\operatorname{argmax}} \sum_{r} \Pr\left(U, r|W, U_1^{t-1}\right) \\ &= \underset{U}{\operatorname{argmax}} \sum_{r} \Pr\left(r|W, U_1^{t-1}\right) \Pr\left(U|W, U_1^{t-1}, r\right) \\ &= \underset{U}{\operatorname{argmax}} \sum_{r} \Pr\left(r|W, U_1^{t-1}\right) \frac{\Pr\left(U|U_1^{t-1}, r\right) \Pr\left(W|U, U_1^{t-1}, r\right)}{\Pr\left(W|U_1^{t-1}, r\right)}\end{aligned} \tag{7}$$

To simplify this expression, we make the same assumptions that we made to obtain (5). Note that we assume that the number of segments r has no influence on the probability of the word sequence or on the probability of the DA. Thus, the new

labelling model is

$$\widehat{U} = \underset{U}{\operatorname{argmax}} \Pr\left(U|W, U_1^{t-1}\right)$$
$$\approx \underset{U}{\operatorname{argmax}} \sum_r \Pr\left(r|W, U_1^{t-1}\right) \sum_{s_1^r} \prod_{k=1}^{r} \Pr\left(u_k|u_{k-(n-1)}^{k-1}\right) \Pr\left(w_{s_{k-1}+1}^{s_k}|u_k\right) \quad (8)$$

As in the previous model, we can obtain a segmentation from (8)

Therefore, we have derived two labelling models from (3). The model described in (5) is the classical approach to dialogue turn labelling. It does not contain any information about the number of segments of the turn nor any information about the segmentation. The model presented in (8) is a new proposal for DA labelling that includes the estimation of the probability of the number of segments.

In (5) and (8), $\Pr(u_k|u_{k-(n-1)}^{k-1})$ 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 Hidden Markov Model (HMM). The estimation of the probability $\Pr(r|W, U_1^{t-1})$ from (8) is explained in the following section. In the implementation of both (5) and (8), the summation over the segmentation and the number of segments is replaced by a maximisation that can be implemented using the Viterbi algorithm.

3 Estimation of the number of segments

In Section 2, we introduced the probability of the number of segments, which we defined as $\Pr(r|W, U_1^{t-1})$. To estimate this probability, first, we consider that the probability of the number of segments does not depend on the previous DAs U_1^{t-1}. Furthermore, we model the dependency of the number of segments r with the output of the speech recogniser W as a function $f(W)$ defined over the sequence of words. Then, the probability of the number of segments can be approximated as

$$\Pr\left(r|W, U_1^{t-1}\right) \approx \Pr(r|f(W)) = \frac{\Pr(f(W)|r)\Pr(r)}{\Pr(f(W))} \quad (9)$$

In this work, we propose two methods to compute this function using the transcription of the turn; however, other approaches could be possible (e.g. using prosodic features obtained from the utterance).

In this proposal, the *a priori* probability $\Pr(r)$ can be easily computed as the number of turns with r segments, N_{Tr}, divided by the total number of turns N_T:

$$\Pr(r) = \frac{N_{Tr}}{N_T} \quad (10)$$

The conditional probability $\Pr(f(W)|r)$ is estimated by a normal distribution. We calculated one distribution for each number of segments r. The mean and variance are computed from the turns with r segments. Finally, $\Pr(f(W))$ is estimated by another Gaussian distribution that is computed from all the turns.

We define two methods to compute $f(W)$ based on some features that can be obtained from the word sequence. We consider that $W = w_1^l$, where l is the number of words in the turn:

Table 1. *Comparison between the two corpora used in the experiments*

	SwitchBoard	Dihana
Dialogues	1,155	900
Turns	115,000	6,280
Vocabulary	42,000	823
Running words	1,837,222	48,243
Task-oriented	No	Yes
Overlaps	Yes	No
Disfluences	Yes	Yes
Interaction	Human-Human	Human-Machine

- *Length of the turn.* We assume that there is a relation between the number of segments and the number of words in a turn. In this case, the function $f(w_1^l)$ can be calculated as the number of words in the turn:

$$f_l(w_1^l) = l \tag{11}$$

- *Final and initial n-grams.* In the transcription, some sequences of words clearly indicate the end of a segment. Moreover, initial sequences can indicate the start of a segment. We propose the computation of $f(w_1^l)$ as the summation of the probability of the sequences of words in the turn being at the end (or at the start) of a segment:

$$f_{fng}(w_1^l) = \sum_{i=n}^{l} \Pr_f(w_{i-(n-1)}^i) \quad f_{ing}(w_1^l) = \sum_{i=n}^{l} \Pr_i(w_{i-(n-1)}^i) \tag{12}$$

  where $\Pr_f(w_{i-(n-1)}^i)$ is the probability of the current n-gram being a final sequence in a segment. This probability is estimated by counting the number of times in the training corpus that the n-gram is at the end of a segment divided by the total number of appearances of the n-gram. This value is 0 for the n-grams that never appear at the end of a segment. Analogously, the probability $\Pr_i(w_{i-(n-1)}^i)$ is referred to initial n-grams. When we use $n = 1$, we are using only the final or initial words.

Obviously, we can obtain methods to estimate the number of segments by combining those two features. It is possible to do a linear combination of the features or to consider that the number of segments depends on different features. We can easily compute the probability of this last case assuming that there are no dependencies among features (naive-Bayes assumption).

4 Corpora

We used two different corpora to test the labelling provided by the two models described in Section 2 ((5) and (8)). These corpora are: SwitchBoard (Godfrey, Holliman and McDaniel 1992) and Dihana (Benedí *et al.* 2006). Table 1 shows the most important features of the two corpora. In the subsections below, we include a complete description of the two corpora.

The SwitchBoard corpus is a well-known corpus in English. It is composed of recorded conversations between humans with no particular goal to accomplish, so it does not represent a real dialogue system. Despite this, SwitchBoard is commonly used to test dialogue labelling methods. The Dihana corpus is a set of conversations in Spanish between a human and a simulated machine. There are several defined scenarios in which the user has to obtain information about train tickets from the system. This corpus was acquired by simulating a real dialogue system and it is task-oriented.

For both corpora we have available the transcription of the turns and a segmented version of the transcription. This version is used to test the model described by (6), which give us the optimistic estimation of the labelling.

4.1 SwitchBoard *Corpus*

The SwitchBoard corpus (Godfrey *et al.* 1992) 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 contains spontaneous speech with frequent interruptions between the speakers and background noises. The transcription of the corpus takes into account all these facts and 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. In average, each turn has 1.8 segments. The vocabulary size is approximately 42,000 words.

The corpus was manually divided into segments following the criteria defined by (Jurafsky, Shriberg and Biasca 1997), and it was annotated using a shallow version of the DAMSL annotation scheme (Core and Allen 1997) known as SWBD-DAMSL. Each segment was 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/nonopinion, yes-no/open/rhetorical-questions, etc.). The manual labelling was performed by eight 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 (Stolcke *et al.* 2000; Webb, Hepple and Wiks 2005).

To simplify the labelling task, we preprocessed the transcriptions of the Switch-Board corpus to remove certain particularities. The interrupted segments were joined to avoid interruptions and ignore overlaps between the speakers. The vocabulary was reduced by using all the words in lowercase. Since we do not have available the SwitchBoard audio, we simulated the output of a speech recogniser by removing the punctuation marks, along with the possible disfluences that could be detected by the speech recogniser. A complete description of the corpus preprocessing can be found in (Martínez-Hinarejos, Granell and Benedí 2006).

To obtain more reliable results, we partitioned 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. The preprocessed corpus and the partitions used in this work are available on the web.[1]

4.2 DIHANA *corpus*

The DIHANA corpus (Benedí *et al.* 2006) is composed of 900 dialogues about a telephone train information system. It was acquired from 225 different speakers (153 male and 72 females), with small dialectal variants. There are 6280 user turns and 9133 system turns. The vocabulary size is 823 words. The total amount of speech signal is about five and a half hours.

The acquisition of the DIHANA corpus was carried out by means of an initial prototype, using the Wizard of Oz (WoZ) technique (Fraser and Gilbert 1991). This acquisition was only restricted at the semantic level (i.e. the acquired dialogues are related to a specific task domain) and was not restricted at the lexical and syntactical levels (spontaneous-speech). In this acquisition process, the semantic control was provided by the definition of scenarios that the user had to accomplish and by the WoZ strategy, which defines the behaviour of the acquisition system.

The annotation scheme used in the corpus is based on the Interchange Format (IF) defined in the C-STAR project (Lavie *et al.* 1997). Although it was defined for a Machine Translation task, it has been adapted to dialogue annotation (Fukada *et al.* 1998). The three-level proposal of the IF format covers the speech act, the concept, and the argument, which makes it appropriate for its use in task-oriented dialogue.

Based on the IF format, a three-level annotation scheme of the DIHANA corpus segments was defined in (Alcácer *et al.* 2005). This DA set represents the general purpose of the segment (first level), as well as more precise semantic information that is specific to each task (second and third levels). The manual labelling was performed by one human labeller, thus the Kappa value is 1.

All of the dialogues are segmented in turns. There are two kinds of turns: those produced by the user and those produced by the system. Each word of the turn has attached a speaker mark (U for user and S for system), and each turn is also divided into segments. Finally, each segment is labelled with a three-level label. Obviously, more than one segment can appear per turn. In fact, an average of 1.5 segments per turn was obtained.

The experiments were performed following a cross-validation approach. The 900 dialogues were divided into five partitions of 180 dialogues. We used only the user turns. To reduce the annotation difficulty, all the dialogues were transcribed in lower case words, and categorised (town names, dates, hours, etc.). The disfluences and punctuation marks were also removed from the transcriptions.

[1] http://users.dsic.upv.es/∼cmartine/research/resources.html

5 Experiments and results

We present a set of five experiments. These 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 ((5) and (8)). The experiments are organised as follows: In Section 5.1 we introduce the evaluation measures used to test the labelling models. In Section 5.2, we compute the baseline experiments of the labelling with the classical HMM-based model (5). In Section 5.3, we show the errors of the estimation of the number of segments with the methods described in Section 3. In Section 5.4, we present the labelling experiments with the new labelling model, which includes the estimation of the number of segments as described in (8). Finally, in Section 5.5, we presented the results of the new labelling model (8) with the real output of a speech recogniser from the DIHANA corpus.

5.1 Evaluation measures

In order to evaluate the labelling models we used the DA Error Rate (DAER) and the Turn Error Rate (TER). The DAER is the average edit distance between the reference DA sequences of the turns and the DA sequences assigned by the labelling model. The TER indicates the percentage of turns that are incorrectly labelled. A turn is incorrectly labelled if the DA sequence of the estimation does not match perfectly the correct sequence of DAs in the turn. For some experiments, we also computed the precision, recall and F-measure as described in (Manning *et al.* 1999).

We also computed a 90 per cent confidence interval for the DAER to ensure statistical significance. This confidence interval was estimated using a bootstrap estimation (Bisani and Ney 2004). Confidence intervals were calculated using bootstrapping with 10,000 repetitions.

5.2 Baseline

In Section 2, we presented two models for labelling. One of these models is the classic approach for turn labelling, which is represented by (5). In this approach, we assume that there is no information about the number of segments in the turn or the segmentation. We also introduced a modification for labelling the turns when a segmentation is available in (6). We consider these two labelling models as the lower and upper baseline for our work. The model described by (5) is the one we want to improve. The model described in (6) is the best labelling we can obtain since it has available the segmentation of the turns. It is a hypothetic case because, in a real system, we do not know the correct segmentation.

Table 2 shows the results of labelling the SWITCHBOARD corpus using 2-gram and 3-gram 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 versions of the corpus. In the segmented version, since we knew the correct segmentation, (6) was used. However, in the unsegmented version, since we did not know anything about the segmentation or the number of segments, (5) was applied. In Table 3, we present the results of the labelling using the same models with the DIHANA corpus.

Table 2. *DAER and TER results with the models described in (5) and (6) for the* SwitchBoard *corpus. The errors are presented for both segmented and unsegmented corpora. The DAER is presented with a 90 per cent confidence interval. The baseline result considered for the next experiments is shown in boldface*

	2-gram	
	DAER	TER
Segmented	31.0 ± 0.2	39.8
Unsegmented	56.2 ± 0.3	55.3
	3-gram	
	DAER	TER
Segmented	31.1 ± 0.2	39.7
Unsegmented	**56.2 ± 0.3**	55.3

Table 3. *DAER and TER results with the model described in (5) and (6) for the* Dihana *corpus. The errors are presented for both segmented and unsegmented corpora. The DAER is presented with a 90 per cent confidence interval. The baseline result considered for the next experiments is shown in boldface*

	2-gram	
	DAER	TER
Segmented	24.2 ± 1.0	24.1
Unsegmented	38.6 ± 1.3	33.1
	3-gram	
	DAER	TER
Segmented	23.8 ± 1.0	23.5
Unsegmented	**37.1 ± 1.2**	32.6

These results are boundary errors and are similar to those provided by (Martínez-Hinarejos *et al.* 2006), where we introduced the HMM model for DA labelling described in (5) and (6). The segmented turns gave us the minimum error supplied by the HMM-based model. The unsegmented turns gave us the maximum error, which was obtained without knowing the segmentation. We consider that the results obtained with the unsegmented version and a 3-gram are baseline errors. Therefore, in SwitchBoard, the baseline DAER is 56.2 per cent. In Dihana, the baseline DAER is 37.1 per cent. These experiments are useful because they allow us to measure the difference between this model and the one with the estimation of the number of segments.

Table 4. *Results of the estimation of the number of segments. The first column indicates the feature used in the estimation of r. The error column indicates the percentage of the turns where the estimated number of segments is different from the real number of segments. It includes the estimation for the* SwitchBoard *and* Dihana *corpora*

	Error	
	SwitchBoard	Dihana
Length	44.7	31.2
Final Words	51.3	46.7
Final Bigrams	**33.6**	**20.1**
Initial Words	48.3	54.9
Initial Bigrams	48.3	28.3

5.3 Estimation of the number of segments

This set of experiments helped us to determine the best way to estimate the number of segments of a turn using the methods introduced in Section 3. Table 4 shows the results of the different estimations of the number of segments for the SwitchBoard and Dihana corpora. The estimated number of segments r is given by: $\hat{r} = \text{argmax}_r \Pr(r|f(W))$. The error measures the percentage of turns where the estimation of the number of segments was wrong.

These tests showed that the final bigrams provided the best estimation of the number of segments for the SwitchBoard corpus. In the Dihana corpus, the final bigrams were the best estimation. In the SwitchBoard corpus, 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 n-grams produced better results due to the presence of certain sequence of words that always indicate the end of a segment (e.g. in the Dihana corpus the dates and destinations; in the SwitchBoard corpus the backchannels words like 'uh-huh'). However, the initial bigrams were good estimators in the Dihana corpus. This difference between corpora may be due to the different nature of the corpora and the fact that the corpora are in different languages. SwitchBoard is composed of human-human dialogues, whereas Dihana is composed of human-machine dialogues that simulate a dialogue system. We also tested the linear combination and the naive-Bayes combination of different features, but the combinations did not produce any significant improvement in the estimation of the number of segments.

5.4 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 (8), where we introduced an estimation of the probability of the number of segments.

For both corpora, we used the estimations of the number of segments tested in subsection 5.3, and 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 5. *DAER and TER results of the labelling of* SwitchBoard *corpus 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. The DAER is presented with a 90 per cent confidence interval. The inclusion of the labelling with the correct r is only for reference*

2-gram		
r estimation	DAER	TER
No estimation	56.2 ± 0.3	55.3
Correct *r*	47.5 ± 0.2	49.1
Length	54.6 ± 0.2	55.1
Final Words	54.2 ± 0.3	54.9
Final Bigrams	**53.3 ± 0.2**	54.1
Initial Words	54.0 ± 0.2	54.7
Initial Bigrams	54.3 ± 0.2	54.9

3-gram		
r estimation	DAER	TER
No estimation	56.2 ± 0.3	55.3
Correct *r*	47.2 ± 0.1	49.1
Length	54.6 ± 0.2	55.1
Final Words	54.2 ± 0.2	54.9
Final Bigrams	**53.2 ± 0.1**	54.0
Initial Words	54.0 ± 0.2	54.7
Initial Bigrams	54.3 ± 0.2	55.1

Table 5 shows a comparison of the errors obtained in the SwitchBoard experiments. The error with correct *r* estimation was computed by labelling the unsegmented corpus, knowing the correct number of segments (i.e. $\Pr(r|f(w))$ 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 a hypothetical case. The rest of the lines refer to different estimations of the number of segments. In Table 6, we present the results of the experiments for the Dihana corpus.

For both corpora, the best result was obtained with the estimation of the number of segments based on final bigrams and the probability of the DA given by a 3-gram. The confidence interval for these experiments and the confidence interval of the baseline errors show that the difference between the results given by the models are statistically significant. Thus, it can be concluded that the model with the estimation of the probability of the number of segments produces a significant improvement in the labelling.

The labelling experiments show that the differences between the estimations of the number of segments are not extrapolated to the labelling process. This is due to two reasons.

Table 6. *DAER and TER results of the labelling of* DIHANA *corpus 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. The DAER is presented with a 90 per cent confidence interval. The inclusion of the labelling with the correct r is only for reference*

2-gram		
r estimation	DAER	TER
No estimation	38.6 ± 1.3	33.1
Correct *r*	26.5 ± 1.0	24.9
Length	36.0 ± 1.3	30.9
Final Words	32.6 ± 1.2	29.1
Final Bigrams	**31.8 ± 1.1**	28.6
Initial Words	34.3 ± 1.2	30.1
Initial Bigrams	32.3 ± 1.2	29.1
3-gram		
r estimation	DAER	TER
No estimation	37.1 ± 1.2	32.6
Correct *r*	25.3 ± 1.0	23.9
Length	35.1 ± 1.2	30.6
Final Words	31.7 ± 1.2	28.5
Final Bigrams	**31.1 ± 1.1**	28.2
Initial Words	33.7 ± 1.2	29.5
Initial Bigrams	31.7 ± 1.2	28.7

First, in the labelling process we do a search over all the segmentations and include an estimation of the probability of the number of segments ($\Pr(r|f(W))$). In the estimation of segments presented in Section 5.3, we only take into account the number of segments r that maximises $\Pr(r|f(W))$. Table 7 shows the estimation of the number of segments produced by the labelling model with 3-gram. In this case, there is a direct relation between the estimation of the number of segments and the DAER.

Second, there are some turns which were not correctly labelled in any of the experiments, even when the correct number of segments was given. As pointed out in (Stolcke *et al.* 2000), the cause of these errors could be that some DA definitions are arbitrary and may even confuse a human labeller. To investigate this problem, we calculated the precision, recall, and F-measure of the experiments.

We present in Table 8 the precision, recall, and F-measure of some experiments with the SWITCHBOARD corpus, and Table 9 shows the results for the DIHANA corpus. The precision indicates the accuracy of the labeller, but the positions of the labels in the labelling are not important; therefore, these errors are better than the corresponding DAER.

Table 7. *Results on the number of segments produced by the labelling model. The first column indicates the feature used in the estimation of r. The error column indicates the percentage of the turns where the estimated number of segments is different from the real number of segments. It includes the estimation for the* SwitchBoard *and* Dihana *corpora using 3-gram in the labelling model*

	Error	
	SwitchBoard	Dihana
No estimation	34.1	18.9
Length	32.8	16.2
Final Words	32.3	11.8
Final Bigrams	32.0	**11.3**
Initial Words	**31.5**	14.7
Initial Bigrams	32.2	13.1

Table 8. *Precision, recall, and F-measure of the labelling of the* SwitchBoard *corpus. 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*

2-gram			
r estimation	Precision	Recall	F-measure
No estimation	0.66	0.47	0.55
Correct *r*	0.60	0.60	0.60
Final Bigrams	0.65	0.51	0.57
3-gram			
r estimation	Precision	Recall	F-measure
No estimation	0.66	0.47	0.55
Correct *r*	0.60	0.60	0.60
Final Bigrams	0.66	0.51	0.57

In the SwitchBoard experiments, the precision is similar for the three experiments compared, which means that the errors are produced by the labeller, even with the correct number of segments. In the Dihana experiments, the results show the improvement produced by the inclusion of the probability of the number of segments in the labelling. However, there are no significant differences in the precision when comparing the experiments with the correct number of segments and the experiments with the estimation based on final bigrams.

These results show that the model with the number of segments produces higher improvements in the Dihana corpus. This is due to the nature of the corpus. Dihana is a corpus with human-machine interaction, simulating a real dialogue system. It

Table 9. *Precision, recall, and F-measure of the labelling of the* DIHANA *corpus. 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*

2-gram			
r estimation	Precision	Recall	F-measure
No estimation	0.69	0.75	0.72
Correct *r*	0.75	0.75	0.75
Final Bigrams	0.72	0.75	0.74
3-gram			
r estimation	Precision	Recall	F-measure
No estimation	0.70	0.76	0.73
Correct *r*	0.76	0.76	0.76
Final Bigrams	0.73	0.76	0.74

was designed for DA labelling that can aid the dialogue manager. The SWITCHBOARD corpus is based on human-human interactions with no specific task to accomplish, and the labelling is ambiguous, as it was pointed out in (Stolcke *et al.* 2000).

5.5 Labelling the speech recogniser output

In the previous sections, we computed the labelling using the transcription of the speech signal. In a real spoken dialogue system, we do not have the correct transcription, so we have to work with the output of a speech recognition software. In this section we present the tests carried out using the two labelling methods ((5) and (8)) with the output of a speech recogniser. We did a experiment with the DIHANA corpus to validate the new labelling model when applying it to a recognised turns. We used the DIHANA corpus because we had available the corpus and the speech recognition system for it.

We tested the models using one of the partitions from the DIHANA corpus, and the other four partitions were used for training. The training data was used to obtain acoustic models (Hidden Markov Models trained with the recorded speech signal) and the language model (a k-TTS automaton (García and Vidal 1990) inferred from the preprocessed transcriptions without punctuation marks). The WER for this test partition was about 20 per cent. The output of the speech recogniser was categorised.

Table 10 includes the results of the baseline model (described by (5)) and the labelling using the new model (8) knowing the correct number of segments *r* and with the estimation of *r* given by the final bigrams. This estimation was used because it produced the best labelling results using the transcribed corpus. We included the labelling errors of the same experiments using the only the first partition of the transcribed version.

Table 10. *DAER and TER results of the labelling of a recognised version of the* DIHANA *corpus 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. The DAER is presented with a 90 per cent confidence interval. The inclusion of the labelling with the correct r is only for reference. The column error for the transcribed corpus shows the labelling error of the transcribed turns of the partition used for speech recognition*

2-gram				
	Recognised turns		Transcribed turns	
r estimation	DAER	TER	DAER	TER
No estimation	37.9 ± 2.9	32.9	41.0 ± 2.9	34.6
Correct *r*	25.2 ± 2.1	25.6	26.2 ± 2.2	25.3
Final Bigrams	31.1 ± 2.5	28.7	32.8 ± 2.5	29.6
3-gram				
	Recognised turns		Transcribed turns	
r estimation	DAER	TER	DAER	TER
No estimation	38.8 ± 3.1	33.0	39.4 ± 2.8	34.9
Correct *r*	25.5 ± 2.2	25.5	25.7 ± 2.1	25.1
Final Bigrams	31.7 ± 2.5	28.9	32.4 ± 2.4	29.8

The results show that the new labelling model is useful to label recognised turns. The inclusion of the estimation of the number of segments in the model does not make the labelling worse. In these experiments the confidence intervals are greater than those obtained in the previous sections. This is due to the smaller size of the test corpus since in these experiments we only labelled one partition.

6 Conclusions and future work

In this work, we have presented 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. Two methods for estimating the probability of the number of segments of a turn based on the transcription are also presented. The new labelling model has been tested with two different corpora: SWITCHBOARD, which is a well-know corpus of human-human conversations; and DIHANA, which is a task-oriented corpus.

The results show that the DA labelling task can be improved by including the probability distribution of the number of segments. Even though our best results

are not as good as the ones obtained using the correct segmentation, they are significantly better than the errors 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. The experiment carried out with the recognised partition of the DIHANA corpus shows that the new labelling model can be used with recognised turns without noticeable degradation in the DA decoding.

Future work is directed towards obtaining new models to estimate the number of segments using new features. In spoken dialogues, a new estimation could be obtained from features that are directly extracted from the audio signal, as proposed in (Ang *et al*. 2005). This new estimation could be included in our probability model of the estimation of the number of segments.

References

Ang, J., Liu, Y., and Shriberg, E. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the International Conference of Acoustics, Speech, and Signal Processings*, vol. 1, pp. 1061–4, Philadelphia.

Alcácer, N., Benedí, J. M., Blat, F., Granell, R., Martínez, C. D., and Torres, F. 2005. Acquisition and labelling of a spontaneous speech dialogue corpus. In *Proceeding of 10th International Conference on Speech and Computer (SPECOM)*. Patras, Greece, pp. 583–6.

Benedí, J. M., Lleida, E., Varona, A., Castro, M. J., Galiano, I., Justo, R., López de Letona, I., and Miguel, A. 2006. Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: Dihana. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, pp. 1636–9.

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

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

Core, M. G., and Allen, J. F. 2007. Coding dialogues with the DAMSL annotation scheme. In *Fall Symposium on Communicative Action in Humans and Machines*. American Association for Artificial Intelligence, pp. 28–35.

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

Fraser, M., and Gilbert, G. 1991. 'Simulating speech systems.' *Computer Speech and Language* (5): 81–9.

Fukada, T., Koll, D., Waibel, A., and Tanigaki, K. 1998. Probabilistic dialogue act extraction for concept based multilingual translation systems. *ICSLP 98* 2771–4.

García, P., and Vidal, E. 1990. Inference of k-testable languages in the strict sense and application to syntactic pattern recognition. *IEEE Transactions on Pattern Analysis Machine Intelligence* **12**(9): 920–5. ISSN: 0162-8828. (1990), IEEE Computer Society.

Godfrey, J., Holliman, E., and McDaniel, J. 1992. Switchboard: telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, pp. 517–20, IEEE.

Gorin, A., Riccardi, G., and Wright, J. 1997. How may I help you? *Speech Communication* **23**: 113–27.

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

Lavie, A., Levin, L., Zhan, P., Taboada, M., Gates, D., Lapata, M. M., Clark, C., Broadhead, M., and Waibel, A. 1997. Expanding the domain of a multi-lingual speech-to-speech translation

system. In *Proceedings of the Workshop on Spoken Language Translation, ACL/EACL-97*, Madrid, Spain.

Levin, L., Ries, K., Thymé-Gobbel, A., and Levie, A. 1999. Tagging of speech acts and dialogue games in Spanish call home. In *Workshop: Towards Standards and Tools for Discourse Tagging*, pp. 42–7.

Manning, C. D., and Sch1/4tze, H. 1999. *Foundations of Statistical Natural Language Processing*, Cambridge, Massachussetts: Massachussetts Institute of Thechnology Press, ISBN:0262133601.

Martínez-Hinarejos, C.-D. 2009. A study of a segmentation technique for dialogue act assignation. In *Proceedings of the Eighth International Conference in Computational Semantics IWCS8*, Tilburg University, Department of Communication and Information Sciences, pp. 299–304.

Martínez-Hinarejos, C. D., Benedí, J. M., and Granell, R. 2008. Statistical framework for a Spanish spoken dialogue corpus. *Speech Communication* **50**: 992–1008.

Martínez-Hinarejos, C. D., Granell, R., and Benedí, J. M. 2006. Segmented and unsegmented dialogue-act annotation with statistical dialogue models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sesions*, Sydney, Australia, pp. 563–70.

Schatzmann, J., Thomson, B., and Young, S. 2007. Statistical user simulation with a hidden agenda. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pp. 273–82.

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

Young, S. 2000. Probabilistic methods in spoken dialogue systems. *Philosophical Trans Royal Society (Series A)* **358**(1769): 1389–402.

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

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