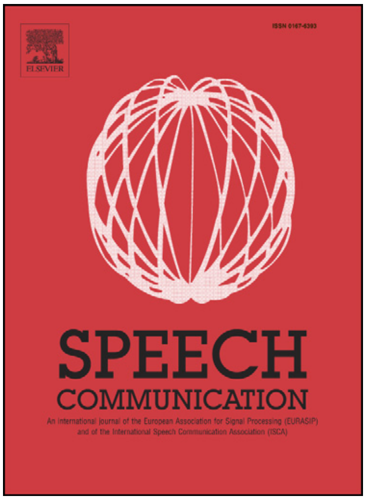
Statistical framework for a Spanish spoken dialogue corpus

Carlos-D. Martínez-Hinarejos, José-Miguel Benedí, Ramón Granell


ELSEVIER
ISSN 0167-6393
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP)
and of the International Speech Communication Association (ISCA)

Statistical framework for a Spanish spoken dialogue corpus

Carlos-D. Martínez-Hinarejos, José-Miguel Benedí, Ramón Granell

PII: S0167-6393(08)00089-7
DOI: 10.1016/j.specom.2008.05.011
Reference: SPECOM 1731

To appear in: *Speech Communication*

Received Date: 12 June 2007
Revised Date: 9 May 2008
Accepted Date: 29 May 2008

Please cite this article as: Martínez-Hinarejos, Carlos-D., Benedí, J.-M., Granell, R., Statistical framework for a Spanish spoken dialogue corpus, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.05.011

Statistical framework for a Spanish spoken dialogue corpus [⋆]

Carlos-D. Martínez-Hinarejos, José-Miguel Benedí [a]
Ramón Granell [b]

[a] *Instituto Tecnológico de Informática, Universidad Politécnica de Valencia, Camino de Vera, s/n, 46022, Valencia, Spain*

[b] *Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford, OX1 3QD, England*

Abstract

Dialogue systems are one of the most interesting applications of speech and language technologies. There have recently been some attempts to build dialogue systems in Spanish, and some corpora have been acquired and annotated. Using these corpora, statistical machine learning methods can be applied to try to solve problems in spoken dialogue systems. In this paper, two statistical models based on the maximum likelihood assumption are presented, and two main applications of these models on a Spanish dialogue corpus are shown: labelling and decoding. The labelling application is useful for annotating new dialogue corpora. The decoding application is useful for implementing dialogue strategies in dialogue systems. Both applications centre on unsegmented dialogue turns. The obtained results show that, although limited, the proposed statistical models are appropriate for these applications.

Key words: Spoken dialogue systems, Statistical models, Dialogue annotation, Dialogue models

1 Introduction

In the last few decades, the development of speech technologies has led to speech-based solutions for several tasks. Dialogue systems are one of the most

[⋆] Work partially supported by VIDI-UPV under project 20070315, by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-C02-01, by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018), and by the EU under project FP6-IST 034434.

challenging examples of those solutions where a computer interacts with a human user to solve a problem or fulfil a task using spoken dialogues (Kuppevelt and Smith (2003)). Since 2004, several important projects involving both academic research groups and high technology companies have been set up in this area, such as CALO (Mark and Perrault (2004)), *Companions* (Wilks (2006)), *Indigo* (Trahanias (2007)), among others. In spite of the fact that one of the objectives of these projects is the development of multimodal, adaptive, human-like spoken dialogue systems for a variety of purposes, they usually concern only a limited knowledge domain. This restriction is mainly motivated by the intrinsic difficulty of the problem and the current state of speech technology. Tasks such as ticket reservation or timetable consultation have usually been considered appropriate for these systems (Aust et al. (1995); Seneff and Polifroni (2000)). This is due to their restricted vocabulary, small set of concepts (which are usually highly structured), and well-defined system tasks.

A dialogue system is traditionally defined as a computer system that can interact with a human being through dialogue in order to complete a specific task (e.g., ticket reservation, timetable consultation, bank operations, etc.). One of the most important aspects of a dialogue system is the capability of modelling the structure of the discourse, i.e., the representation of the current state of the dialogue and the definition of what the dialogue system must do at each point.

A profound understanding of the discourse structure is a multidisciplinary problem that does not currently have a clear solution. According to (Schiffrin (1994)), there are six types of approaches to discourse analysis: interactional sociolinguistics, conversation analysis, the ethnography of communication, variation analysis, pragmatics, and speech act theory. This last discourse theory (Austin (1962)) focuses on communicative acts performed through speech and is the framework in which many authors try to model the structure of dialogue. From a more practical point of view and partially based on this theoretical approach, specific solutions have been proposed to model discourse in dialogue problems using a wide range of methods: dialogue grammars (McTear et al. (2000)), information state (Bos et al. (2003)) and the reinforcement learning framework (Williams and Young (2007)), among others. Independent of the method, many of these proposals make use of Dialogue Acts (DA) to model the local structure of the dialogue. In this paper, we are interested in this first level of analysis, which involves the detection of DAs. A DA represents the meaning of an utterance, where the utterance can be defined as a dialogue-relevant subsequence of words in the current user turn (Aust et al. (1995)).

Similar to other NLP problems (like speech recognition and understanding, or statistical machine translation), data-based approaches to dialogue modelling such as Stolcke et al. (2000) and Young (2000) have been developed in the last decade. These machine learning approaches rely on statistical models that can be automatically estimated from annotated data, which in this case, are dialogues from the task (knowledge domain).

In statistical modelling, the appropriate parameters of the models are learnt from examples. In statistical discourse modelling, the examples are annotated dialogues. As a simplification, each situation in the dialogue can be associated with a specific label, and the models learn how to identify and react to the different situations by estimating the associations between the labels and the dialogue events (words, previous turns, etc.). Therefore, annotation schemes based on DA definitions must be defined to annotate the dialogues and to infer the statistical model parameters.

Several DA annotation schemes have been proposed in recent years: Dialogue Act Mark-up in Several Layers -DAMSL (Core and Allen (1997)), *VerbMobil* (Alexandersson et al. (1998)), DATE (Walker and Passonneau (2001)), and Dihana (Alcácer et al. (2005)), among others. From a practical point of view, in almost every dialogue project, a DA annotation scheme is selected according to the nature of the dialogue system and the approach taken to the problem. Usually, it is common to reuse an existing scheme and adapt it to the specific features of the corpus. For example, SWBD-DAMSL (Jurafsky et al. (1997)) and ICSI-MRDA (Shriberg et al. (2004)) are variations of the DAMSL scheme. In all these studies, it is necessary to annotate a large number of dialogues in order to estimate the parameters of the statistical models. Manual annotation is the usual solution, although it is very time-consuming and there is a tendency for error, since the annotation instructions are not easy to interpret or apply, and human annotators can commit errors (Jurafsky et al. (1997)).

Therefore, the application of semi-automatic annotation techniques is of significant interest. The *labelling* process performed by these techniques consists of automatically annotating every dialogue turn in one or more DA according to the annotation scheme. Different knowledge sources from the corpus can be used to develop this task, such as transcribed words, speech features, turn order, etc. This labelling process can be stated as finding the most likely DA sequence for the given dialogue by making use of statistical models. Many recent works have attempted this approach for different purposes. Levin et al. (1999) try to find the most likely speech act sequence from the words of an utterance for the purpose of predicting more abstract labels called dialogue games. In Stolcke et al. (2000), the purpose of labelling is to improve the speech recognition performance of the utterance, and it is applied to segmented turns. Both of them make use of a combination of N-grams and Hidden Markov models, but they report experiments over different corpora (CallHome Spanish and SwitchBoard), obtaining 56% and 70% accuracy in DA labelling, respectively. Webb et al. (2005) make use of word N-grams of utterances for DA classification on segmented turns. With this simple approach, they report results similar to Stolcke et al. (2000) for the same corpus. Also with the same data, a relative improvement of 12% in performance is obtained by Rangarajan et al. (2007) using maximum entropy modelling with prosodic, lexical and syntactic features.

All these works, except Levin et al. (1999), rely on the availability of segmentation of the turns into subsequences of words that correspond to DAs (utterances). However, turn segmentation into utterances is not commonly available. Other works rely on a decoupled scheme of segmentation and DA classification (Ang et al. (2005)).

Another interesting application is immediately derived from the labelling process. When implementing a dialogue system, the dialogue strategy usually defines the reaction of the system to each situation. The dialogue strategy takes into account several variables that are available at the current point of the dialogue. One of the most important variables is the last user turn intention, which is reflected in the DA sequence associated to that turn. To estimate the DA associated to a user turn, statistical models can be used. This so-called *decoding* process differs from the labelling application mainly due to the lack of available information in every dialogue turn since it is an on-line process. Most of the works reported above did not perform this task.

In this paper, we present two statistical annotation models that compute both the labelling (off-line process) and decoding (on-line process) of dialogue turns. Our models perform in the more realistic situation where the segmentation of turns is not available, although they can be easily adapted to segmented turns. The models are based on types of models that have shown good performance in other language processing tasks (such as automatic speech recognition or machine translation). The models were previously used in preliminary works that revealed some features and limitations. For this paper, we made a more in-depth analysis and systematic usage of the models. A comparative analysis of the labelling performance of the two models was carried out. Their application to the decoding problem is also reported. The goal was to determine their robustness and the degradation of quality that is produced with lower information sources (including the lack of segmentation of the turns), along with the appropriateness of the models to spoken dialogue systems. This entire evaluation process was made on a spoken Spanish dialogue corpus.

Spontaneous-speech dialogue corpora in Spanish are not very common; therefore, opportunities to test dialogue technologies on Spanish data are also rare. In this article, the corpus selected for testing the proposed models is the Dihana *corpus* (Benedí et al. (2006)), which is a task-oriented telephone spontaneous-speech dialogue corpus in Spanish. The Dihana task consists of the retrieval of information about the Spanish national train network by telephone. In order to limit the domain, the queries are restricted to timetables and fares for long-distance, nationwide trains (Benedí et al. (2006)).

This paper is organised as follows: Section 2 presents the statistical models; Section 3 describes the dialogue corpus used in the experiments; Section 4 establishes the experimental framework and presents a summary of the results; Section 5 presents our conclusions and future research directions.

2 Statistical Dialogue Act Modelling

The statistical framework used in this paper is described as follows: given a word sequence $\mathcal{W}$ obtained from the recognition module of a spoken dialogue system, the main goal is to obtain the optimum DA sequence $\widehat{\mathcal{U}}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$, i.e., we seek:

$$\widehat{\mathcal{U}} = \arg\max_{\mathcal{U}} \Pr(\mathcal{U}|\mathcal{W}) \tag{1}$$

Our main interest is to solve this maximisation for an unsegmented word sequence. In order to estimate this posterior probability, we consider two alternatives: the application of Bayes' rule to express the posterior probability in terms of much more straightforward models (Stolcke et al. (2000),Young (2000)) or the application of transducers learnt by Grammatical Inference techniques (Vidal (1994)). These alternatives are reflected in two different models that provide two different methods of solving the optimisation problem of expression (1).

The first model is the HMM-based model, which is based on the application of Bayes' rule. The HMM-based model is inspired in the classical use of Hidden Markov Models (HMM) and N-grams in Automatic Speech Recognition. In our case, the words of the turn are the emitted features and the DA labels act as the identifiers of the HMM models that emit the words. A language model is required to determine the probability of the DA sequences, and weight parameters should be used to balance the influence of the emitting and the language model. This type of model has previously been used to automatically annotate dialogues with DA, but the reported experiments assumed the availability of the segmentation of dialogue turns into utterances (Stolcke et al. (2000)). Our proposal is to apply this type of model on unsegmented dialogue turns. Previous experiments using this type of model on unsegmented turns (Martínez-Hinarejos et al. (2006)) showed a moderate recognition accuracy but a poor segmentation accuracy. This fact makes us think that this model will provide limited results when used on unsegmented dialogue turns.

The second model is the NGT model, which implements the posterior probability directly. The NGT model is inspired on a Machine Translation technique that is based on the inference of Finite State Transducers. In our case, the input sentence is formed by the sequence of words of the turn, and the output sentence is the sequence of DA labels that result from the translation. Previous experiments using this model (Martínez-Hinarejos (2006)) showed a moderate recognition accuracy and a good segmentation accuracy, although the experiments were performed under different conditions from those of Martínez-Hinarejos et al. (2006). Therefore, this model could be more suitable when used on unsegmented dialogue turns.

The two models have different features, parameters and estimation processes. Our interest is to apply these two models on unsegmented dialogue turns, although both models can be easily adapted to the segmented case, which would allow them to be compared with the results provided in previous works (Levin et al. (1998); Stolcke et al. (2000); Webb et al. (2005)).

2.1 HMM-based Model

Expression (1) maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$. Applying Bayes' rule, we get:

$$\widehat{\mathcal{U}} = \arg\max_{\mathcal{U}} \Pr(\mathcal{U}|\mathcal{W}) = \arg\max_{\mathcal{U}} \Pr(\mathcal{U}) \Pr(\mathcal{W}|\mathcal{U}) \tag{2}$$

where $\Pr(\mathcal{U})$ represents the prior probability of a DA sequence, and $\Pr(\mathcal{W}|\mathcal{U})$ is the likelihood of the word sequence. Following the work of other authors (Stolcke et al. (2000); Young (2000)), estimating $\Pr(\mathcal{U})$ requires the definition of a statistic model of DA sequences, typically N-grams, while the HMM are usually used to model how the input word sequence is produced given the underlying DA sequence.

Expression (2) allows us to obtain the best DA decoding $\widehat{\mathcal{U}}$ of the complete dialogue. This expression is very useful for dealing with the problem of off-line labelling of a dialogue corpus. Although segmentation of the turns into utterances is not usual in transcribed dialogues, many previous works on this problem assume the availability of the segmentation (Stolcke et al. (2000)). Therefore, to allow a comparison with previous work, two variants will be considered: with segmentation available, and without segmentation available.

However, these assumptions cannot be considered in an on-line dialogue system. In this case, the only information available is the information previous to the current user interaction, along with the information given with the current user interaction (i.e., look-ahead is not possible). The usual information associated to the user interaction is: transcribed word sequences (for text systems), recognised word sequences, and phonetic/prosodic features (for speech systems). Moreover, the segmentation of the last user interaction into utterances is usually not available. Therefore, it will be necessary to develop a model to cope with both the segmentation and the decoding problem for the current user interaction.

Given the DA sequence $\mathcal{U}$ of the complete dialogue, $U_1^{t-1} = U_1 U_2 \cdots U_{t-1}$ represents the DA sequence detected until the current turn t. Let $W = w_1 w_2 \ldots w_l$ be the word sequence of the current turn, where l is the sequence length. In this work, we have only used the information of the word sequence.

Given W, the word sequence of the current turn, we can describe W in terms of a possible segmentation as: $W = W_1^l = W_{s_0+1}^{s_1} W_{s_1+1}^{s_2} \ldots W_{s_{r-1}+1}^{s_r}$; where r is the number of segments, $s = (s_0, s_1, \ldots, s_r)$ is the segment representation, and s_k is the index of the segment k of W. It is important to note that r is bounded and only can take values between 1 and l.

From here, we can reformulate expression (1) introducing a new posterior probability $\Pr(U|W_1^l, U_1^{t-1})$; where U is the probability of the DAs sequence associated to the current user turn given the word sequence of the current user turn W_1^l and the history of previous DA sequences U_1^{t-1}. Applying Bayes' rule again, we can rewrite expression (1), as

$$\widehat{U} = \arg\max_U \Pr(U|W_1^l, U_1^{t-1}) = \arg\max_U \Pr(U|U_1^{t-1}) \Pr(W_1^l|U, U_1^{t-1}) \qquad (3)$$

Then, we rewrite both probability distributions by the introduction of the 'hidden' segmentation $s = (s_0, s_1, \ldots, s_r)$ and the number of segments r. Therefore, U can be expressed as $U = u_1^r$, and W as $W_1^l = W_{s_0+1}^{s_1} W_{s_1+1}^{s_2} \ldots W_{s_{r-1}+1}^{s_r}$. In order to structure these probability distributions, we factorise them over the position in the segment:

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

To simplify these models, we introduce some assumptions: 1) The probability of the word segments is independent of both the history of previous DA sequences U_1^{t-1} and the previous DA sequences of the current user turn u_1^{k-1}; and 2) the probability of DAs only depends on the n previous DAs. Therefore, by replacing these restrictions in expression (3) and substituting the summation by a maximisation on r and s_1^r, the result is

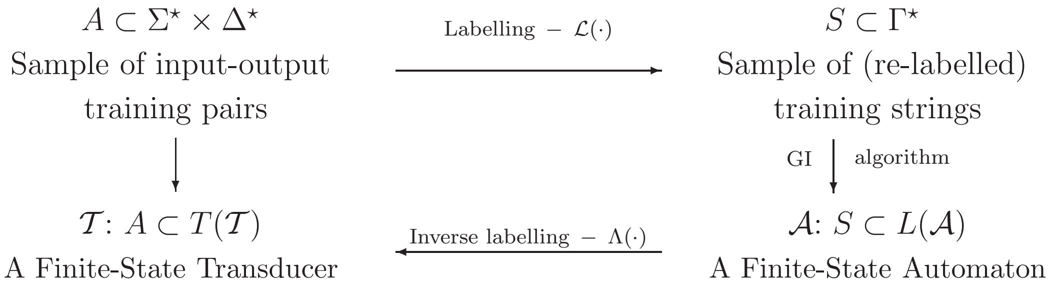

$A \subset \Sigma^\star \times \Delta^\star$
Sample of input-output training pairs
Labelling − $\mathcal{L}(\cdot)$
$S \subset \Gamma^\star$
Sample of (re-labelled) training strings
GI algorithm
$\mathcal{T}$: $A \subset T(\mathcal{T})$
A Finite-State Transducer
Inverse labelling − $\Lambda(\cdot)$
$\mathcal{A}$: $S \subset L(\mathcal{A})$
A Finite-State Automaton

Fig. 1. General scheme for the GIATI technique. Σ, Δ and Γ are the input, output, and extended set of symbols, respectively. A and S are the initial sets of aligned and re-labelled samples. $L(\mathcal{A})$ and $T(\mathcal{T})$ represent the languages derived from $\mathcal{A}$ and $\mathcal{T}$, respectively. $\mathcal{L}$ and Λ are the labelling and inverse labelling functions.

$$\widehat{U} = \arg\max_{U} \max_{r,s_1^r} \prod_{k=1}^{r} \Pr(u_k|u_{k-n-1}^{k-1}) \Pr(W_{s_{k-1}+1}^{s_k}|u_k) \qquad (4)$$

These models can be easily implemented using simple statistical models (N-grams and Hidden Markov Models). The maximisation (including the segmentation and the DA decoding) can be implemented using the Viterbi algorithm. A grammar scale factor, which is similar to the one used in speech recognition, can be incorporated into the model to control the weight of the language model ($\Pr(u_k|u_{k-n-1}^{k-1})$) in the Viterbi process.

Expression (4) can be directly applied on the labelling problem with unsegmented turns. In the case of segmented turns, the model skips the maximisation over (r, s_1^r), since s_1^r is given. In the decoding problem, expression (4) is applied turn by turn, and no changes in the decoding results of previous turns are allowed (i.e., the previously assigned labels are fixed). In Section 4.2.1, we present some experiments that are related to the labelling problem associated with expression (4). In Section 4.3.1, we also present some experiments that are related to the decoding problem associated with expression (4).

2.2 N-Gram Transducer Model

The next proposal is the N-Gram Transducer (NGT) model. The objective is to avoid having to use models of a different nature, which must be estimated with different techniques and combined using weight factors to obtain a final model. The NGT model directly estimates the posterior probability of expression (1) by means of a transducer. Thus, in this case there is no need to infer and combine models of a different nature (as happened with the previously presented HMM-based model), and no weight factors need to be used or tuned to obtain the optimal results.

The definition of this model is based on a Stochastic Finite-State Transducer (SFST) inference technique known as GIATI[1] (Casacuberta et al. (2005)). GIATI is a general SFST inference technique, which is based on a re-labelling process of input-output pairs of sentences. This re-labelling process depends on the alignments between the input and the output symbols (Brown et al. (1993)). A grammatical model is inferred from the re-labelled corpus. This model is transformed into the final SFST by inverting the re-labelling process.

The general scheme for the GIATI technique is presented in Figure 1. In the first step, a re-labelling process (which is the key point in this technique) is applied over the input-output training pairs, building a re-labelled sample. In the second step, this re-labelled sample is used to infer a Stochastic Finite-State Automaton (SFSA) with a grammar inference algorithm. In the last step, the re-labelling process is inverted on the inferred SFSA to derive the final SFST.

The specific application of GIATI to the dialogue problem is much easier than its application to the general translation problem. In this application, the input symbols are the words of the turn, and the output symbols are the associated DA. The alignment between input and output pairs is defined by aligning the last word of each utterance with the corresponding DA and leaving the rest of the words in the turn with an empty alignment. With this alignment strategy, the resulting alignment is linear, i.e., there are no cross-inverted alignments. This facilitates the selection of the re-labelling scheme. It should also be pointed out that this technique is independent of the language and the application that the dialogue covers. In any case, both the words of the language and the DA labels are treated as generic sets of symbols, which makes the application of the technique independent of these factors. A preliminary application of GIATI to dialogue annotation that covers these points was presented in Martínez-Hinarejos and Casacuberta (2000).

In this work, the re-labelling scheme (first step) is as follows:

- If a word w is not aligned with any DA, then the new label is the same word w.
- If a word w is aligned with a DA d, then the new label is $w@d$, where @ is a special joining metasymbol that is not present in Σ (input language) nor in Δ (output language).

Figure 2 shows an example of re-labelling with this scheme for a simple sample sentence ("Yes , from Madrid ."). In this example, the last word of each utterance (",", ".") is re-labelled into a new word composed of that word plus the joining symbol and the DA label (",@Acceptance", ".@Answer"). The rest

[1] GIATI is the acronym for Grammatical Inference and Alignments for Transducer Inference.

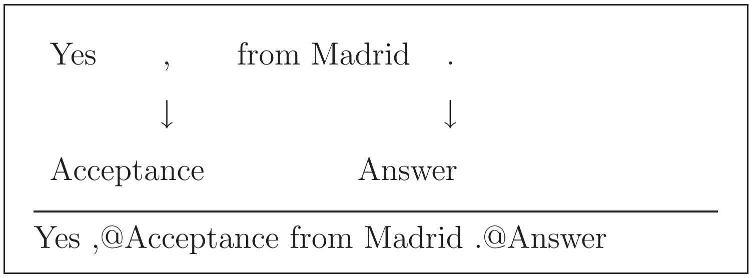


Fig. 2. An example of re-labelling for a turn of the task. The upper part shows the alignment between the words and the DA, and the lower part shows the result of the re-labelling.

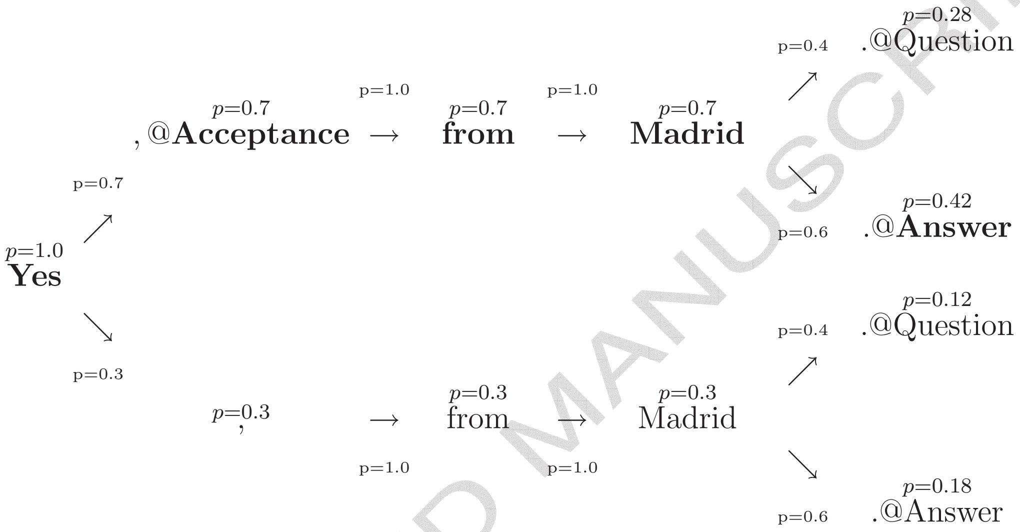


Fig. 3. An example of the Viterbi tree search, along with the evolution of the probabilities. The branch that results in the maximum probability is shown in boldface.

of the words are re-labelled as the same words.

The second step is the inference process. In our case, this process computes a smoothed N-gram, where N is a parameter. This allows the technique to take advantage of the well-established smoothing techniques for N-grams, which have successfully demonstrated their appropriateness for language modelling. The resulting N-gram computes the probabilities of all the sequences of N words in the re-labelled training data (i.e., from the sequence in the lower part of Figure 2). An equivalent SFSA can be computed from this smoothed N-gram.

The third step is the transformation of the SFSA into a SFST by applying the inversion of the re-labelling. The SFST can process an unlabelled dialogue and provide the corresponding DA labels for each utterance.

The conversion of a smoothed N-gram into a SFSA is difficult because there is no clear or efficient mechanism to implement back-off transitions in a SFSA.

In our case, the SFSA/SFST transformation is avoided by implementing a Viterbi algorithm that works directly with smoothed N-grams using them as transducers.

This Viterbi implementation takes the words in the current turn as input and performs a tree beam-search. The ith level in the tree corresponds to the sequence of the first i words of the turn. Each node in the tree corresponds to a sequence of words and its corresponding outputs, along with the probability of the sequence. Therefore, when taking the next word, all the nodes in the previous level branch into k child nodes, where k is the number of outputs that the current word has associated to it. The probability of a child node is calculated from the probability of its parent and the probability of the new N-gram sequence (which results from concatenating the parent's N-gram with the current word and its output). In the last level, the node with the highest probability is chosen, and the final output is obtained by going up the tree and retrieving the corresponding sequence of words and outputs. To limit the search space, beam-search was implemented in this Viterbi exploration in order to control the size of the tree expansion.

Figure 3 shows an example of tree expansion for the dialogue turn shown in Figure 2. The N-gram is inferred from a set of sentences of the task and the expansion is done for the input "Yes , from Madrid .". The algorithm processes "Yes", which only has one alternative, and calculates the probability for this node. Then, it processes ",", which has two alternatives ("," and ",@Acceptance"). Then, it branches the search into two nodes; the probability for each of them is calculated from the probability of their parent node and the N-gram probability of "Yes ," and "Yes ,@Acceptance", respectively. Figure 3 shows the N-gram probability next to the arrows. This process continues until the end of the input. Then the node with the highest probability is used to retrieve its corresponding branch (in Figure 3, this node is the second node in the last column), which gives the final output sequence.

This model is clearly a highly local model, that is, it only takes into account local information (the context of n words) to make the assignation of labels. This high locality allows the model to be more robust to global errors in the input turns (this is important if the input turns are a result of a speech recognition process, which may be error-prone). However, this high locality prevents the model from taking into account information on the dialogue structure at a higher level (e.g., the dialogue history) in an appropriate way. This fact can be critical in the obtaining of coherent DA sequences, as the influence of previous turns gets lost in most cases because only the last $n-1$ words are taken into account to choose the optimal DA label. Some previous works showed that when a model does not use this information, it can negatively affect its performance (Stolcke et al. (2000)); however, other works (Webb et al. (2005); Martínez-Hinarejos et al. (2006)) showed that the influence of the dialogue

Table 1
Features of some Spanish spontaneous-speech dialogue corpora. Where (#D) is the number of dialogues; (#V) is the vocabulary size; (S) are the dialogue actors: (Human and System); and (Task) is a brief description of the task.

Corpus	#D	#V	S	Task
Vestel DB	16,000	3,000	H-S	Questions about personal data, y/n questions and spelling words
NIST	742	-	H-H	Cross-channel conversations about 70 different topics
CallHome	120	13,000	H-H	Family conversations without topic restriction
Restaurant	523	-	H-H	Retrieving information about menu items from a fast food restaurant

history is not so critical.

This model can be applied to the labelling problem on segmented and unsegmented dialogues by appropriately restricting the output on the search process in each step. We consider both the segmented and unsegmented variants in order to compare our results with previous works. Experiments on labelling with this model are presented in Section 4.2.2. Its application to the decoding problem is also possible by fixing the search branches previous to the current turn. Experiments on decoding with this model are presented in Section 4.3.2.

3 Spanish Spoken Dialogue Corpus

Even though Spanish speech language resources are not as common as English ones, it is possible to find some corpora of very different types and with several dialectical variations. However, almost all of these resources are of read speech. Only very few of them are of spontaneous-speech, and even fewer involve dialogue (Tapias et al. (1994); Finke et al. (1998); Cieri (2004); López-Cózar et al. (1998)). Table 1 shows four of the spontaneous-speech dialogue corpora that are available in Spanish. All of these corpora, except the one described in López-Cózar et al. (1998), are non-task-oriented.

The dialogue corpus used in this work was a spontaneous-speech dialogue corpus in Spanish: Dihana corpus (Benedí et al. (2006)). Dihana corpus is composed of 900 task-oriented computer-to-human spoken dialogues. In order to limit the task domain, the queries were restricted to timetables, fares, and services for long-distance trains in Spanish.

Comparing the Dihana corpus with other corpora of the same nature (task-oriented dialogues using spontaneous-speech), Dihana can be considered as a medium-sized corpus according to the number of dialogues and as a small corpus according to the vocabulary size (823 words). Table 2 shows some basic features of various task-oriented dialogue corpora used in other projects. These corpora vary widely in size, from a couple of dozen dialogues to several

Table 2
Features of some task-oriented dialogue corpora. Where (#D) is the number of dialogues; (#V) is the vocabulary size; (#W) are the running words; (#S-T) is number of speaker turns; (S) are the dialogue actors: (Human and System); (L) is the Spoken language: (English, German, Italian, Japanese and Spanish); and (Task) is a brief description of the task.

Corpus	**#D**	**#V**	**#W**	**#S-T**	**S**	**L**	**Task**
Taba	5,200	2,545	90,377	33,568	H-S	G	Retrieving information about train timetables
Arise-IT	1,909	3,475	-	-	H-S	I	Retrieving information about train times and fares
MERCURY	1,700	1,586	-	25,000	H-S	E	Flight reservations
Dihana	900	823	48,243	15,413	H-S	S	Retrieving information about train times and fares
Verbmobil G-VM1	793	7,120	308,028	30,750	H-H	G	Scheduling appointments
Communicator travel	648	-	314,223	11,715	H-S	E	Travel planning
Verbmobil J-VM2	220	4,447	165,755	12,897	H-H	J	Preparing a business trip
Maptask	128	1,675	150,000	21,251	H-H	E	Cooperative game to draw a route
The Sundial	100	1,500	-	-	H-S	E	Retrieving information about flight times and fares
TRAINS 93	98	860	55,000	5,900	H-H	E	Manufacturing and shipping goods in a railroad freight system
The Monroe	20	1,550	52,000	4,794	H-H	E	Coordinating solutions in emergency scenarios

thousand. Trains93 corpus (Heeman and Allen (1994)) might be considered as the corpus that is the most similar to Dihana due to its vocabulary size and task. However, an equivalent task is also performed in the Communicator travel corpus (Walker et al. (2001)) but with a larger vocabulary than Dihana. Vocabulary size, which is an indicator of the complexity of the task, not only depends on the size of the corpus (words, turns and dialogues), but also on other features such as language and the nature of the speakers who perform the dialogues (usually human-human dialogues have more variation than system-human dialogues). Another feature used to determine the difficulty of the task is the semantic concepts that are defined in the dialogue system. In Dihana, there are 13 concepts (origin, destination, departure time, arrival time, day, fare, duration, train type, trip type, number of trains, class, services, relative number of train); however, in most of the dialogues, not more than five concepts were needed to complete the prefixed scenario during the acquisition.

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 level (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 (Fraser and Gilbert (1991)), which defines the behaviour of the acquisition system.

The Dihana corpus was acquired from 225 different speakers (153 male and 72 female), with small dialectal variants. These 900 dialogues comprise 6,280 user turns and 9,133 system turns. On average, each dialogue consisted of seven user turns and ten system turns, with an average of 7.7 words per user turn. The vocabulary size was 823 words. The total amount of speech signal was about five and a half hours.

The different turns were segmented into utterances. Obviously more than one utterance can appear per turn. In fact, an average of 1.5 utterances per turn was obtained. Each utterance was identified with a DA and was annotated with a DA label.

Before describing the DA labelling process, it is necessary to define the DA set. This is an important issue since this DA label set should cover many aspects of the user interactions, such as intention, provided and required data, general aim of the utterances, etc. One of the most commonly used DA sets is the Dialogue Act Markup in Several Layers (damsl) set (Core and Allen (1997)) employed in projects such as Amitiés (Hardy (2002)). This set was defined for the annotation of human-to-human collaborative dialogues in complex tasks. For this reason, the DA set covers complex communicative and interactive functions, which are usually not in most of the less sophisticated task-oriented dialogue systems devoted to information systems. Therefore, some variations of the damsl scheme have been used mainly for human-to-human task-free dialogue corpora, such as SwitchBoard (using an adapted set known as swbd-damsl and keeping only the more frequent labels, Jurafsky et al. (1997)); ICSI meeting corpus (where the tagset was adapted for using in multiparty dialogues Shriberg et al. (2004)); and CallHome Spanish (using the clarity set, Levin et al. (1998)). However, some task-oriented dialogue corpora, such as TRAINS (Heeman and Allen (1994)) or Amitiés (Hardy (2002)), used damsl as annotation scheme, but they rely on internal data state models to keep track of the informational issues concerning the task. In our case, we are interested in an annotation scheme that can keep both the basic communicative functions and the data flow.

Several other schemes have been proposed in the last ten years. mate (Mengel et al. (2000)), which is based on swbd-damsl, and the MapTask DA set (Carletta et al. (1996)) are examples of popular DA sets. Nevertheless, they were devised for human-to-human dialogues and complex tasks and are not adequate for annotating the Dihana corpus.

One well-suited scheme is the Interchange Format (IF) defined in the C-STAR project (Lavie et al. (1997); Reaves et al. (1998)). Although this interlingua scheme was defined for a Machine Translation task, it has been applied 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 use in task-oriented dialogues.

Based on the IF format, a three-level annotation scheme of the Dihana corpus utterances was defined in Alcácer et al. (2005). This DA set represents the general purpose of the utterance (first level), as well as more precise semantic information that is specific to each task (second and third levels). The second level contains the repository of information implicit in the utterance (i.e., the set of data used or modified according to the intention given by the first level). The third level represents the specific data present in the utterance. The DA set used for each level is presented in Table 3. All of the dialogues are segmented in turns (User and System), and each turn is also segmented into utterances. Finally, each utterance is labelled with a three-level label.

All the dialogues were manually transcribed. These transcriptions were used to annotate the corpus by means of a semiautomatic procedure (Alcácer et al. (2005)). Next, all the dialogues were manually corrected by human experts using a very specific set of defined rules (Alcácer et al. (2005)). The annotation of all the dialogues was consistently revised by a single expert. Figure 4 shows a sample of annotated dialogue (in English) from the Dihana corpus. After this process, there were 248 different labels (153 for user turns, 95 for system turns) using the three-level scheme. When considering only the first and second levels, there were 72 labels (45 for user turns, 27 for system turns). When considering only the first level, there were only 16 labels (7 for user turns, 9 for system turns). The coverage of the DA labels on the corpus does not favour the use of a very small set of labels. For example, with the three-level scheme, 60 labels (22% of the total, 34 for user, 26 for system) cover 90% of the utterances, and with the two-level scheme, 22 labels (30% of the total, 9 for user, 13 for system) cover 90% of the utterances.

Before using this corpus, some automatic preprocessing was performed to reduce the complexity of the corpus and the structures. This is necessary in order to obtain more robust models because data in its raw form has a high variability and sparseness. The preprocessing included the following points:

- A categorisation process was performed for categories such as town names, the time, dates, train types, etc.
- All the words were transcribed in lowercase.
- Punctuation marks were separated from words.
- All the words were speaker-labelled (U for user, S for system).

Speaker	Utterance	Transcription		
		Level 1	Level 2	Level 3
S	S1	Welcome to the railway information system. How may I help you?		
		Opening	Nil	Nil
U	U1	I want to know the departure times from Valencia		
		Question	Departure-hour	Origin
	U2	to Madrid		
		Question	Departure-hour	Destination
	U3	arriving on May the 15th of 2,004.		
		Question	Departure-hour	Day
S	S2	Do you want to leave on Saturday, May the 15th of 2,004?		
		Confirmation	Day	Day
U	U4	Yes.		
		Acceptance	Day	Nil
S	S3	Consulting times for trains from Valencia to Madrid on Saturday, May 15th of 2,004.		
		Confirmation	Departure-hour	Destination, Day, Origin
	S4	Wait a moment, please.		
		Waiting	Nil	Nil
	S5	There are several trains. The first one leaves at 7:45 and arrives at 11:14, and the last one leaves at 18:45 and arrives at 22:18.		
		Answer	Departure-hour	Arrival-hour, Departure-hour Order-number, Number-trains
	S6	Do you need anything else?		
		Consult	Nil	Nil
U	U5	Yes, I want to know the fare for the train leaving at 7:45.		
		Question	Fare	Departure-hour
S	S7	That train in tourist class costs 35.50 euros.		
		Answer	Fare	Class, Fare
	S8	Do you need anything else?		
		Consult	Nil	Nil
U	U6	No, thank you.		
		Closing	Nil	Nil
S	S9	Thanks for using this service. Have a nice day.		
		Closing	Nil	Nil

Fig. 4. An example of an annotated dialogue in English from the Dihana corpus. *Nil* denotes the absence of information.

In general, the Dihana corpus can be viewed as a medium-sized, spontaneous-speech dialogue corpus involving a well-known task. Its features show Dihana as one of the first and largest task-oriented corpora in Spanish. All the dialogues were transcribed and annotated. The DA annotation scheme follows many of the principles used in other projects with a three-level structure that covers the general intention as well as more specific details of the domain of the task.

Table 3
DA labels defined for the three levels and an example for each one from the Dihana corpus.

First level	**Example**
Opening	Welcome to the railway information system. How may I help you?
Closing	No, thank you for the information.
Undefined	I hate machines like you, what do you have to say?
Not-understood	I'm sorry, I didn't understand you. Can you repeat that?
Waiting	Wait a moment, please.
Consult	Do you need anything else?
Acceptance	Yes, please.
Rejection	No!
Second level	**Example**
Departure-hour	I want to know the times for trains leaving Madrid.
Arrival-hour	Do you want to arrive before 8:00?
Fare	Yes, I want to know the fares.
Origin	Departing from Zaragoza?
Destination	What town do you want to go to?
Day	On March the 15th.
Train-type	Do you want to travel on Euromed?
Service	Can I get my car on the train?
Class	Do you want to travel in business class?
Trip-time	The trip takes 4 hours and 34 minutes on Altaria.
Third level	**Example**
Departure-hour	I want to know the times before ten.
Arrival-hour	Yes, times and fares for the one arriving before eight.
Fare	I want a cheap train.
Origin	I want to go from Madrid to Toledo.
Destination	I want to go from Madrid to Toledo.
Day	On next Thursday.
Train-type	The fare on Talgo in tourist class is 10.50 euros.
Service	Non-smokers, please.
Class	How much is travel in business class?
Trip-time	The trip lasts for 3 hours and 12 minutes.
Order-number	What is the fare for the first train?
Number-trains	There is only one train, departing at 7:35.
Trip-type	Yes, the return times.

4 Experiments and Results

In this section, we report the experiments that were carried out with the Dihana corpus to test the performance of the two models (HMM-based and NGT) described in Section 2. The aim of the experiments was to confirm the appropriateness of the two models on unsegmented input and to evaluate their possible complementarity. To evaluate the behaviour of these models, we present experimental results for two different tasks:

Labelling: This refers to obtaining all the DA labels associated to a complete dialogue (i.e., off-line process, look-ahead is allowed). These experiments are needed to assess the quality of the models in obtaining correct annotations of dialogues. Labelling is important for obtaining fast and accurate annotation of corpora, which can be used to infer more accurate models. Although our main interest is the use of labelling when the segmentation is not available, we also performed experiments with the segmentation available. These last results act as an upper-bound in performance and allow a comparison with previous works by other authors (Levin et al. (1998); Stolcke et al. (2000); Webb et al. (2005)).

Decoding: This refers to obtaining the DA labels of a spoken turn. In this case, only the previous turns of the dialogue are available (i.e., on-line process, look-ahead is not allowed). Moreover, it is assumed that the segmentation is not available. These experiments assess the quality of the models in a real dialogue system implementation, which requires a correct interpretation of the user interactions in order to obtain an appropriate response from the system. The experiments have been designed to consider perfect recognition (decoding on the transcription) and speech recognition (decoding on the recognition).

4.1 Initial experiments

Our proposal is not the first one that has attempted to solve the labelling task defined above. There have been multiple works on dialogue labelling on several corpora using different statistical models. Most of these have used segmented data (Levin et al. (1998); Stolcke et al. (2000); Webb et al. (2005)), while a few have used unsegmented turns (Warnke et al. (1997); Ang et al. (2005)). Although the exhaustive experiments were done with the Dihana corpus, we also performed some previous experiments on another corpus to be able to evaluate the performance of our models and to compare our results with those of other authors.

To do this, we decided to use another well-known Spanish corpus, the CallHome Spanish corpus (Levin et al. (1998)). CallHome Spanish is a database of 120 unscripted (i.e., non-task-oriented) telephone dialogues between native speakers of Latin-American Spanish on the telephone line. The topics and word choice were completely unrestricted. The conversations were between relatives, which leads to a very informal speaking style and very spontaneous speech. All the dialogues were transcribed with special word labels for noises and non-linguistic events (laughs, coughs,. . .). The vocabulary is composed of approximately 13,000 words. The DA label set used for its annotation was the Clarity set (Finke et al. (1998)).

Some experiments were performed using conditions similar to those presented in (Levin et al. (1998)). A total of 80 dialogues were used for training and 40 dialogues were used for testing, and the annotation consisted of 63 different labels. The labelling process was applied to the segmented dialogues in order to make a fair comparison of the results. For the HMM-based model under optimal conditions, an accuracy of 69% was obtained. For the NGT model a slightly lower accuracy was obtained (66%). These results are similar to those reported by other authors (in Levin et al. (1998), an accuracy of 69% is reported).

The nature of this corpus (human-to-human, non-task-oriented, with a large vocabulary) is quite different from the nature of the Dihana corpus, and therefore the correlation between the results with CallHome Spanish and Dihana may not be significant. However, these results reveal that our models are appropriate for the labelling task in this corpus.

4.2 Labelling Results

The first application of the models is the labelling of dialogues. In other words, the models are applied over an unlabelled dialogue, and they provide a hypothesis on the labels for each turn, taking into account all the information that is available. Both models were used in this case, and two different variants were proposed:

(1) Segmentation available: This is the optimistic bound of the results and is used as a reference to compare with other authors' results (Stolcke et al. (2000)). In the case of the HMM-based model, the probability of each HMM was computed for each utterance, and the N-gram probabilities were used to compute the probabilities of the transition from one utterance to the next one. In the case of the NGT model, the tree expansion was limited to force no output in the non-final words of the utterances, and to force output in the last word of the utterance.

(2) No segmentation available: This is a more realistic and interesting experiment because it is usually difficult to get a corpus that is utterance-segmented. In this case, the only segmentation available was the one provided by the turns. Therefore, in the HMM-based model, an unrestricted search in the search space was implemented. In the NGT model, the tree expansion was unrestricted (outputs could be emitted in any word). The only exception was that the output was forced in the changes of turns.

Table 4
Dihana corpus statistics (average of the five cross-validation partitions).

	Training			Test		
	User	System	Total	User	System	Total
Dialogues	720			180		
Turns	5,024	7,206	12,330	1,256	1,827	3,083
Utterances	7,773	11,064	18,837	1,406	2,765	4,171
Running words	42,806	119,807	162,613	10,815	29,950	40,765
Vocabulary	762	208	832	417	174	485

Dihana is a medium-sized corpus, and the statistical significance of the results could be biased for the selection of the training and test partitions. Therefore, to obtain significant results in the labelling task with the Dihana corpus, a cross-validation approach was adopted and 5 different partitions were used. Each of them had 720 dialogues for training and 180 for testing. The statistics for the corpus are presented in Table 4. The transcriptions of the dialogues were used for the labelling experiments.

To verify the influence of the number of DA, the results were obtained by considering the complete labels (with first, second and third levels) or the labels with only the first two levels. In the following, they are denoted as three- and two-level labels, respectively. Different utterances were considered in each case because consecutive utterances that only differed in the third level were joined to obtain the two-level segmentation.

4.2.1 HMM-based model labelling results

The first experiment was the annotation with the HMM-based model having both the complete dialogue and the segmentation available. This corresponds to the implementation of expression (4) presented in Section 2.1 with s_1^r given. The experiments were performed using the cross-validation approach and the manually transcribed dialogues. Different weights for the language model were tested, and we show the results for 10000 and 50000, which offered the best results.

The evaluation was done using several measures. In the case of the segmented experiment, errors could be computed at the utterance level or at the turn level. At the utterance level, the error was computed as the number of mislabelled segments with respect to the total number of segments. The results are presented in Table 5, and they are somewhat equivalent to the results

Table 5
Labelling results for the HMM-based model with cross-validation using *segmented* dialogues. Error rate at the utterance level for all the turns.

		2 levels			3 levels		
N-gram		2	3	4	2	3	4
Weight	10000	6.9	6.6	6.6	10.8	10.3	10.5
	50000	7.3	7.1	7.5	14.1	14.4	15.3

Table 6
Labelling results for the HMM-based model with cross-validation using *segmented* dialogues. DAER at the turn level for all the turns.

		2 levels			3 levels		
N-gram		2	3	4	2	3	4
Weight	10000	7.0	6.6	6.6	10.8	10.3	10.5
	50000	7.3	7.1	7.5	14.1	14.4	15.3

obtained by the application of the models presented in Stolcke et al. (2000).

At the turn level, DA Error Rate (DAER) was the computed error measure; DAER corresponds to the classical Word Error Rate (WER) definition but applied to DA labels. The DAER results are presented in Table 6, and they allow us to compare these results with the results when segmentation is not available.

The results in this case reveal an excellent behaviour of the models. The baseline classifier assigns the most frequent DA label to each utterance, and this produces a chance error rate which is more than 80% in the two-level case, and more than 85% in the three-level case. The best error rate with this model was lower than 7% at the two-level labels and around 10% at the three-level labels, which dramatically improved the chance error rate.

The next set of experiments was applied to unsegmented dialogues. In this case, *precision* and *recall* measures can be computed. Precision is computed as the number of correct labels with respect to the hypothesised labels, and recall is computed as the number of correct labels with respect to the reference labels. These measures have the same value as the error rate when the segmentation is available (because the number of hypotheses is equal to the number of references); however, in this case (since insertions and deletions may occur), they provide more information.

The labelling obtained in this experiment was analysed using the DAER mea-

Table 7
Labelling results for the HMM-based model with cross-validation using *unsegmented* dialogues. DAER at the turn level for all the turns.

		2 levels			3 levels		
N-gram		2	3	4	2	3	4
Weight	10000	10.6	9.4	9.3	17.8	17.0	17.2
	50000	11.4	11.2	12.3	18.2	19.3	20.0

Table 8
Precision, recall and F-value results for the HMM-based model with cross-validation using *unsegmented* dialogues at the turn level for all the turns.

N-gram	2			3			4		
	2 levels								
Weight	Prec	Rec	F	Prec	Rec	F	Prec	Rec	F
10000	89.9	92.1	91.0	91.2	92.9	92.0	91.2	93.0	92.1
50000	89.6	91.4	90.5	89.9	91.9	90.9	89.0	91.3	90.2
	3 levels								
10000	83.9	88.7	86.3	84.7	89.5	87.0	84.6	89.3	86.9
50000	86.2	84.9	85.6	84.9	84.9	84.9	84.4	84.0	84.2

sure. Based on previous work with this models, we can expect a moderate to high relative increment of the error rate. The results are presented in Table 7. The precision and recall measures (and the corresponding F-values) are presented in Table 8. Firstly, the results are highly correlated with the DAER measure, which reveals that taking DAER as the quality measure in this case is as appropriate as taking precision and recall.

As expected, the performance of the model degraded with the absence of segmentation. The error rate at two-level labels went up to 9% (50% of the relative increase). The error rate at three-level labels went up to 17% (70% of the relative increase). In any case, the error rates were still much better than the chance error rate (more than 82% at the two-level labels and more than 88% at the three-level labels). These increases reveal the importance of having an adequate segmentation of the turns, which is in consonance with the results reported in Martínez-Hinarejos et al. (2006).

Table 9
Labelling results for N-gram Transducers with cross-validation using *segmented* dialogues. Error rates at the utterance level for all the turns.

N-gram	2	3	4	5
2 levels	28.1	13.6	11.7	11.8
3 levels	36.7	23.4	21.6	21.5

Table 10
Labelling results for N-gram Transducers with cross-validation using *segmented* dialogues. DAER at the turn level for all the turns.

N-gram	2	3	4	5
2 levels	28.1	13.6	11.7	11.8
3 levels	36.7	23.4	21.6	21.5

4.2.2 NGT model labelling results

Similar experiments were carried out with the NGT model, which directly implements the general optimisation problem presented in expression (1). The first experiment again used complete dialogues with segmentation, with the described cross-validation approach and transcribed dialogues. The N-gram degrees applied in the inference process were from two to five. In this case, since there was only one model, there were no extra parameters to tune (whereas in the HMM-based model the N-gram weight had to be tuned).

The same evaluation measures were applied in this case. For segmented turns, the errors at the segment level (number of misannotated segments with respect to the total number of segments) are presented in Table 9 and the errors at the turn level (DAER results) are presented in Table 10.

The first conclusion we can draw from these results is the high influence of the N-gram degree. Optimal results were obtained with 4-grams, which is in consonance with the usual behaviour of the N-gram models in language modelling (the performance increases until a certain degree and, after this threshold, data sparseness causes lower performance). The other main conclusion is that, even in optimal conditions, this technique performs worse than the HMM-based model (5% to 10% of error above the optimal results with the HMM-based model, a relative increase of around 80% for two-level labels and over 100% for three-level labels). This could be caused by the data requirements of the different models, since HMM (which model words) and N-grams (which model labels) individually require less data to adequately estimate their parameters; however, an N-gram that models words and labels at the same time would require a higher amount of data to achieve a similar performance. Moreover, the

Table 11
Labelling results for N-gram Transducers with cross-validation using *unsegmented* dialogues. DAER at the turn level for all the turns.

N-gram	2	3	4	5
2 levels	30.1	15.2	13.3	13.6
3 levels	38.5	25.0	23.0	23.1

Table 12
Precision, recall and F-value results for N-gram Transducers with cross-validation using *unsegmented* dialogues at the turn level for all the turns.

	2 levels			3 levels		
N-gram	Prec	Rec	F	Prec	Rec	F
2	66.9	49.4	56.9	64.4	42.1	51.0
3	84.9	62.8	72.2	83.3	54.6	66.1
4	87.1	64.4	74.1	85.2	55.8	67.6
5	86.8	64.2	73.9	84.8	55.5	67.2

high locality of this type of model (in contrast with the HMM-based model) reveals that the availability of high-level dialogue information (i.e., dialogue history) is necessary to achieve better results.

The next experiment was performed with complete, non-segmented dialogues. The aim was to evaluate the sensitivity of this technique to the lack of correct segmentation (which should be lower than the sensitivity presented by the HMM-based model, according to the better segmentation capacity of the NGT model). Both DAER and precision-recall measures were computed. The DAER results are presented in Table 11. Precision and recall results are presented in Table 12.

In this case, DAER reveals that, although the error rate increases, the relative increase is quite a bit lower than the increase that the HMM-based model presents. In the best case, for two-levels labels the relative error increase was about 14%, and for three-level labels, it was about 6% (in the HMM-based model, they were 50% and 70%, respectively). These results reveal a lower sensitivity of this technique to the availability of the correct segmentation, although the absolute results are still worse than those produced by the HMM-based model (4% to 6% higher error rates). More interesting conclusions can be obtained from the analysis of the precision and recall results. They reveal that recall is quite a bit lower with respect to the value obtained for the HMM-based model. Therefore, the main conclusion is that the NGT model tends to

Fig. 5. Error frequency for the user turn errors for the HMM-based model and the NGT model. The abscissa axis indicates the specific errors. The errors are ordered with respect to the difference of absolute occurrence in each model.

emit a lower number of labels by turn, which is clearly detrimental for the annotation of turns with more than one utterance. Turns of this type are 35% of the total number of turns when using two-level labels, and 44% when using three-level labels. This fact explains the higher error rate with this model. In this case, the better segmentation capacity of the NGT model with respect to the HMM-based model did not have sufficient beneficial influence.

4.2.3 Error analysis

A more in-depth error analysis was performed to check the nature of the errors produced by the two models. An initial analysis revealed that the nature of the errors was different depending on the model used. For example, in the best conditions (segmented and complete dialogues, using 4-grams, and for two-level labels), the HMM-based model tends to change the quantity of slots of the second level, or to change the second-level label, while the first-level label is usually correct. For the same conditions, in the case of the NGT model, the most frequent error is the assignation of the most frequent label, followed by labels that are incorrect at the first level.

To show the difference between the errors, the difference in the percent of error due to each mistake was computed for both models. Figure 5 shows these results for two-level labels and 4-grams. Each point of the abscissa axis represents a specific DA confusion. The order of these errors was chosen using the difference of the frequency of absolute errors. Only user turn errors that were common to the two models were taken. Therefore, the first point in the curves shows an error that occurs simultaneously with a high frequency with the NGT model and with a low frequency with the HMM-based model. Consequently, the last point shows an error that occurs with a high frequency with the HMM-based model and with a low frequency with the NGT model. The middle points reflect the errors that appear in a similar number in the two models. This curve shows that, in general, the most frequent errors with the HMM-based model usually present a lower frequency with the NGT model, and vice-versa. There are a few exceptions for errors that are equally frequent with both models.

4.3 Decoding Results

The next task consists of using our models in the DA decoding of a turn in order to provide the dialogue system with the essential information of the last interaction with the user. The correct detection of DA is critical to obtain an appropriate performance of the dialogue system because the current DA sequence is a fundamental part of most dialogue strategies.

In this case, the only available information is the information previous to the current turn (i.e., the rest of the dialogue is not available). This situation is the most common in spoken dialogue systems. Both the HMM-based model and the NGT model can be used for this task, and two different experimental situations are proposed: perfect decoding (i.e., using manual transcriptions) and speech decoding (i.e., using recogniser output, which may contain errors). In any case, the segmentation is not available.

The HMM-based model implements the decoding by obtaining the DA labels *turn by turn*, i.e., for each turn, a sequence of DA labels is obtained as a hypothesis, and this hypothesis is not changed when more turns are analysed. The NGT model implements the decoding by obtaining the DA labels of the current turn, keeping the optimal branch and erasing the rest of the branches. By this method, the final optimal branch is computed by keeping the decisions made in each turn.

The experiments were carried out on only one of the cross-validation partitions, since that partition was used to perform a speech recognition experiment. The transcriptions of this partition were used for the perfect decoding experiment, and the results from the speech recognition experiment were used for the speech decoding experiment. In the speech recognition experiment, 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%.

In these experiments, the DA detection is only evaluated for user turns, assuming a correct response of the system in all the interactions (i.e., the results are lower bounds of errors). Only DAER results are reported, since the precision and recall results presented in Subsections 4.2.1 and 4.2.2 showed a high correlation with DAER measures.

Table 13
Decoding results for the HMM-based model for the speech test partition using *manual transcriptions*. DAER at the turn level for all the user turns.

		2 levels			3 levels		
N-gram		2	3	4	2	3	4
Weight	10000	22.6	21.8	22.3	39.9	38.7	38.4
	50000	23.9	22.9	24.2	38.1	39.4	39.4

Table 14
Decoding results for HMM-based model with *cross-validation* partitions using *manual transcriptions*. DAER at the turn level for all the user turns.

		2 levels			3 levels		
N-gram		2	3	4	2	3	4
Weight	10000	23.8	23.4	23.2	37.9	36.3	36.9
	50000	24.1	23.6	24.9	37.6	38.0	38.1

4.3.1 HMM-based model decoding results

In this case, the implemented model is the one that corresponds to expression (4) presented in Section 2.1. The first experiment was performed on the test partitions, using manual transcription. The decoding process used each turn (without segmentation) as input. The process was applied for both two- and three-level labels, and the same language model weights (10000 and 50000) were used in the tests. The results are presented in Table 13.

To verify whether or not the chosen partition is representative of the complete corpus, the same experiment was performed with all the cross-validation partitions. The obtained results (Table 14) show no significant differences, and therefore, the chosen partition can be considered as representative.

The first conclusion that can be extracted from these results is the high error rate with respect to the error rates in the labelling application (see comparative results in Table 7 with respect to results in Table 14). This is due to the higher variability and syntactically spontaneous nature of the user turns. Actually, the labelling results over user turns offer a similar error rate, as detailed in Table 15 for reference only (the comparison with results in Table 14 reveals minimum differences). In any case, these error rates are quite high, but better than chance error rates (80% for two-level labels and 85% for three-level labels).

This comparison reveals that the availability of the complete dialogue is not

Table 15
Labelling results (user turns) for HMM-based model using *unsegmented* dialogues. DAER at the turn level.

		2 levels			3 levels		
N-gram		2	3	4	2	3	4
Weight	10000	24.5	22.0	21.6	37.3	35.9	36.2
	50000	23.0	21.6	22.6	36.2	37.9	38.8

Table 16
Decoding results for the HMM-based model for the speech test partition using *recogniser output*. DAER at the turn level for all the user turns.

		2 levels			3 levels		
N-gram		2	3	4	2	3	4
Weight	10000	31.5	30.4	30.3	49.9	50.8	50.0
	50000	27.8	27.0	27.2	40.8	43.9	43.5

critical to obtain optimal assignation of DA (the relative increase is about 7% for two-level labels and about 1% for three-level labels). Thus, the HMM-based model is robust enough to be applied for DA decoding.

The next experiment was to obtain the DA decoding on the recogniser outputs. The results are presented in Table 16.

These results allow us to draw a few conclusions. The first one is the increase in the error rate when using the recognised sentences (as expected). This is clearly significant when the best results for both transcription results (Table 13) and recognition results (Table 16) are taken: the relative increase is around 25% for two-level labels, and around 7% for three-level labels. This is coherent with the WER that the recognition of the test corpus achieved (20%). Another important conclusion is that the language model weight should be higher to obtain optimal results. This is sound with the conditions because in the training of the models, the words (transcribed) differ with respect to those of the test data (recognised), but the dialogue structure is the same. Therefore, the N-gram language model is more accurate for these conditions than the HMM models, and the higher influence of the language model provides better results.

4.3.2 NGT model decoding results

The same experiments were performed with the NGT model. The implementation of expression (1) in this case allows the turn by turn decoding. Initially,

Table 17
Decoding results for the NGT model for the speech test partition using *manual transcriptions*. DAER at the turn level for all the user turns.

N-gram	2	3	4	5
2 levels	55.3	43.9	39.9	41.0
3 levels	60.5	52.0	48.8	49.1

Table 18
Decoding results for NGT with *cross-validation* partitions. DAER at the turn level for all the user turns.

N-gram	2	3	4	5
2 levels	54.5	42.0	39.9	40.1
3 levels	60.1	51.3	49.5	48.7

Table 19
Labelling results (user turns) for N-gram Transducers using *unsegmented* dialogues. DAER at the turn level.

N-gram	2	3	4	5
2 levels	51.1	37.9	33.0	34.3
3 levels	59.3	49.1	45.1	45.8

the experiment was performed with manual transcriptions of the turns. The decoding was performed turn by turn, using both two- and three-level labels. The results are presented in Table 17. To verify the representativity of the partition, the cross-validation experiments were performed, and their results (Table 18) showed no significant differences.

Again, the high increase in the DAER due to the evaluation of only the user turns is relevant. When comparing these results with the results provided for the labelling application (results in Table 19), it is clear that the increase in the error rate is quite significant (in the best results, the relative increase is 20% for two-level labels, and 8% for three-level labels). The high error rates in Table 19 (for user turns only) with respect to the results in Table 11 again reflect the difficulty of modelling user turns. Therefore, the availability of the complete dialogue is much more critical in the case of NGT than in HMM.

Moreover, as occurred in the labelling case, the results of the NGT model are worse than those of the HMM-based model. In this case, the relative increase of the error rates is around 35% for two-level labels and around 70% for three-level labels, which are lower than those reported for the labelling process.

Table 20
Decoding results for the NGT model for the speech test partition using *recogniser output*. DAER at the turn level for all the user turns.

N-gram	2	3	4	5
2 levels	67.6	58.3	60.5	61.5
3 levels	75.6	68.9	67.5	66.4

Fig. 6. Error frequency for the user turn errors in the decoding task for the HMM-based model and the NGT model. The abscissa axis indicates the specific errors. The errors are ordered with respect to the difference of absolute occurrence in each model.

The DAER results for the recognised turns are presented in Table 20. These results show that the NGT model is quite sensitive to the presence of speech misrecognitions because the relative error increase (compared with the results in Table 17) is quite high (50% for two-level labels and nearly 40% for three-level labels). This can be explained by the nature of the NGT model, which needs the accurate recognition of the n last words of the utterance to provide an accurate label (i.e., it uses local information, whereas HMM uses all the information of the utterance). Therefore, misrecognitions in this part of the utterance are critical to the behaviour of the model.

4.3.3 Error analysis

Error analysis was performed for the decoding results, using a methodology that is very similar to the one employed for the labelling results. The analysed results correspond to two-level labels and 3-grams. Again, the nature of the errors is different. The HMM-based model usually produces substitution errors where the second level changes in the number of slots; however, in general, the first level is well recognised. The NGT model has a greater tendency to insertion errors and to change the first level. The curves in Figure 6 reveal the same behaviour as those in Figure 5, and similar conclusions can be inferred.

5 Conclusions and Future Work

In this work, we have presented a statistical framework that can be used for common tasks of dialogue systems. This statistical framework is based on the maximum likelihood approach that assigns DA to sequences of words. Two different models were developed: one based on the Bayes' rule using well-known and straightforward models (HMM and N-grams), and another one based on the direct implementation of a model of the *a posteriori* probability (transducer).

These statistical models were used for two major aims: the labelling of dialogues and the decoding of turns. The labelling process is needed to improve the productivity of the annotation process, which in turn is needed to infer statistical models from corpora. The decoding process is needed to implement dialogue strategies based on the decoding of the user interactions, turn is one of the most important variables to take into account when choosing the system reaction. In both tasks, our main interest centred on the application of the models on unsegmented dialogue turns, which is in contrast with many previous approximations which assumed the segmentation of the turns.

Initial experiments were developed on the CallHome Spanish corpus in order to assess our models and compare our results with other authors. A telephone spontaneous-speech dialogue corpus in Spanish (Dihana corpus) was used for most of the experiments. This corpus was manually annotated. Speech recognition results of a test partition of this corpus were obtained. The statistical models were applied to this corpus to obtain the experimental results in both the labelling and the decoding processes.

The results obtained in this study demonstrate that statistical models perform better than the baseline classifier, even though they still present high error rates. More research is needed to improve them. Our results with the CallHome Spanish corpus are similar to those reported by other authors. The results on unsegmented transcribed dialogues show that statistical models can speed-up the annotation of dialogue corpora. The decoding results show that statistical models are an appropriate starting point for improving the dialogue management.

In general, the HMM-based model performed better than the NGT model. The HMM-based model is significantly independent of the availability of the complete dialogue, which makes it ideal for the decoding task. However, it is quite sensitive to recognition errors and to the lack of segmentation. It also requires an extra parameter (the N-gram weight) to be tuned.

The NGT model had poorer results (due mainly to the low output rate). It is extremely sensitive to the lack of the complete dialogue and to recognition errors. However, it does not need extra parameters to be tuned and is less sensitive to the lack of segmentation.

Future work will be aimed at solving some of the drawbacks of the models and at evaluating them on other corpora and languages. Another line of research is the application of other models (such as belief networks (Meng et al. (2003))) on the presented tasks. It would be interesting to elaborate quality segmentation models that could be applied before the application of the HMM-based model. In the case of the NGT model, some parameters should be added to control the output rate and increase the quality of the results, especially to make high-level dialogue information available for the model. One possibility is to change the labelling step of the GIATI inference process by adding the DA history to the words of the utterance.

The error analysis revealed the different nature of the errors that the two models produce. Therefore, the combination of these two models using the combination of classifiers paradigm (Ho et al. (1994)) is another interesting approach. Finally, as these models are independent from the language and the set of labels, more experiments should be performed with other dialogue corpora in order to generalise the conclusions presented here.

References

Alcácer, N., Benedí, J., Blat, F., Granell, R., Martínez, C. D., 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. 583–586.

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B., Siegel, M., Jul. 1998. Dialogue acts in VERBMOBIL-2 (second edition). Tech. Rep. 226, DFKI GmbH, Saarbrücken, Germany.

Ang, J., Liu, Y., Shriberg, E., 2005. Automatic dialog act segmentation and classification in multiparty meetings. In: Proceedings of the International Conference of Acoustics, Speech, and Signal Processing. Vol. 1. Philadelphia, pp. 1061–1064.

Aust, H., Oerder, M., Seide, F., Steinbiss, V., 1995. The philips automatic train timetable information system. Speech Communication 17, 249–263.

Austin J.L., 1962. How to Do Things with Words. London, Oxford University Press.

Benedí, J.-M., Lleida, E., Varona, A., Castro, M.-J., Galiano, I., Justo, R., López de Letona, I., Miguel, A., May 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. 1636–1639.

Bos, J., Klein, E., Lemon, O. and Oka, T., 2003. DIPPER: Description and Formalisation of an Information-State Update Dialogue System Architecture. 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Mercer, R. L., 1993. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics 19 (2), 263–311.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G. and Anderson, A., 1996. HCRC dialogue structure coding manual. Human Communication Research Centre, University of Edinburgh.

Casacuberta, F., Vidal, E., Picó, D., 2005. Inference of finite-state transducers from regular languages. Pattern Recognition 38 (9), 1431–1443.

Cieri Ch., 2004. The Mixer Corpus of Multilingual, Multichannel Speaker Recognition Data.

Core, M. G., Allen, J. F., Nov. 1997. Coding dialogues with the DAMSL annotation scheme. In: Traum, D. (Ed.), Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines. American Association for Artificial Intelligence, Menlo Park, California, pp. 28–35.

Finke, M., Lapata, M., Lavie, A., Levin, L., Mayfield, L., Tomokiyo, Polzin, T., Ries, K., Waibel, A. and Zechner, K., 1998. CLARITY: Inferring Discourse Structure from Speech. Applying Machine Learning to Discourse Processing (AAAI'98).

Fraser, M., Gilbert, G., 1991. Simulating speech systems. Computer Speech and Language 5, 81–99.

Fukada, T., Koll, D., Waibel, A. and Tanigaki, K., 1998. Probabilistic Dialogue Act Extraction for Concept Based Multilingual Translation Systems. ICSLP 98, pages 2771–2774. Sydney, Australia.

García, P., Vidal, E., 1990. Inference of k-testable languages in the strict sense and application to syntactic pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell. 12 (9), 920–925.

Hardy, H., Baker, K., Devillers, L., Lamel L., Rosset S., Strzalkowski T., Ursu C. and Webb N. 2002. Multi-Layer Dialogue Annotation for Automated Multilingual Customer Service. Proceedings of the ISLE Workshop on Dialogue Tagging for Multi-Modal Human Computer Interaction, Edinburgh, Scotland.

Heeman, P. and Allen, J. 1994, The TRAINS93 dialogues. Technical Report TRAINS TN 94-2, U. Rochester

Ho, T. K., Hull, J. J., Srihari, S. N., January 1994. Decision combination in multiple classifier systems. IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1), 66–75.

Jurafsky, D., Shriberg, E., Biasca, D., 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual - draft 13. Tech. Rep. 97-01, University of Colorado Institute of Cognitive Science.

Kuppevelt, J. V., Smith, R. W., 2003. Current and New Directions in Discourse and Dialogue. Vol. 22 of Text, Speech and Language Technology. Springer.

Lavie A., Levin, L., Zhan, P., Taboada, M., Gates, D., M. Lapata, M., Clark, C., Broadhead, M. and Waibel, A., 1997. Expanding the Domain of a Multilingual Speech-to-Speech Translation System. 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. Proceedings of the Workshop: Towards Standards and Tools for Discourse Tagging, pp. 42–47.

Levin, L., Thymé-Gobbel, A., Ries, K., Levie, A. and Woszczyna, M., 1998. A discourse coding scheme for conversational spanish. ICSLP'98.

López-Cózar, R., Rubio, A.J., García, and P., Segura, J.C., 1998. A spoken dialogue system based on a dialogue corpus analysis. Proc. of LREC, pp. 55-58

Mark, B., Perrault, R., 2004. CALO: a cognitive agent that learns and organizes. <http://www.ai.sri.com/project/CALO>.

Martínez-Hinarejos, C. D., 2006. Automatic annotation of dialogues using n-grams. In Petr Sojka, Ivan Kopecek, and Karel Pala, editors, Proceedings of the Ninth International Conference on Text, Speech and Dialogue–TSD 2006, Lecture Notes in Artificial Intelligence LNCS/LNAI 4188, pages 653-660, Brno, Czech Republic, Sep 2006. Springer-Verlag.

Martínez-Hinarejos, C. D., Casacuberta, F., Sep. 2000. A pattern recognition approach to dialog labelling by using finite-state transducers. In: Proceedings of 5th. IberoAmerican Symposium on Pattern Recognition. Lisbon, Portugal, pp. 669–677.

Martínez-Hinarejos, C. D., Granell, R., 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, 17th-21th July, pp. 563–570.

McTear, M., Allen, S., Clatworthy, L., Ellison, N., Lavelle, C. and McCaffery, H., 2000. Integrating Flexibility into a Structured Dialogue Model: Some Design Considerations. Proc 6th International Conference on Spoken Language Processing, Beijing, China (1). 110-113.

Meng, H. M., Wai, C., Pieraccini, R., 2003. The use of belief networks for mixed-initiative dialog modeling. IEEE Transactions on Speech and Audio Processing 11 (6), 757–773.

Mengel, A., Dybkjaer, L., Garrido, J.M., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A. and Soria, C., 2000. MATE: Dialogue Annotation Guidelines, January. <http://www.ims.uni.stuttgart.de/projekte/mate/mdag/>.

Rangarajan, V., Bangalore, S. and Narayanan, S., 2007. Exploiting prosodic features for dialog act tagging in a discriminative modeling framework. Interspeech, Antwerp, Belgium.

Reaves, B., and Nishino, A. and Takezawa, T., 1998. ATR-MATRIX: Implementation of a speech translation system. Proc. Acoust. Soc. Japan Spring Meeting. pp. 53-54.

Schiffrin D., 1994. Approaches to Discourse. Blackwell textbooks in linguistics. ISBN 0-631-16622.

Seneff, S., Polifroni, J., 2000. Dialogue management in mercury flight reservation system. In: ANLP-NAACL. pp. 1–6.

Shriberg, E., Dhillon, E., Bhagat, S., Ang, J. and Carvey, H., 2004. The ICSI meeting recorder dialog act (MRDA) corpus. Proc. 5th SIGdial Workshop on Discourse and Dialogue. pp. 97–100.

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

Tapias, D., Acero, A., Esteve, J. and Torrecilla, J.C., 1994. The VESTEL Telephone Speech Database. Proc 3th International Conference on Spoken Language Processing Yokohama, Japan. pp. 1811-1814.

Trahanias, P., 2007. Indigo: Interaction with Personality and Dialogue Enabled Robots. <http://www.ics.forth.gr/indigo/>.

Vidal, E., 1994. Grammatical Inference: An Introductory Survey. Proc. of 2nd Grammatical Inference and Applications. LNAI, Vol.862.

Walker, M., Aberdeen, J., Boland, J., Bratt, E., Garofolo, J., Hirschman, L., Le, A., Lee, S., Narayanan, S., Papineni, K., Pellom, B., Polifroni, J., Potamianos, A., Prabhu, P., Rudnicky, A., Sanders, G., Seneff, S., Stallard, D., and Whittaker S., Darpa communicator dialog travel planning systems: The june 2000 Collection. In EUROSPEECH 2001.

Walker, M. and Passonneau, R., 2001. DATE: a dialogue act tagging scheme for evaluation of spoken dialogue systems. HLT'01: Proceedings of the first international conference on Human language technology research, San Diego. pp. 1–8.

Warnke, V., Kompe, R., Niemann, H., Nöth, E., 1997. Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models. In: Proc. European Conf. on Speech Communication and Technology. Vol. 1. Rhodes, pp. 207–210.

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

Wilks, Y., 2006. COMPANIONS: Intelligent, Persistent, Personalised Interfaces to the Internet. <http://www.companions-project.org>.

Williams, J. and Young, S., 2007. Partially observable Markov decision processes for spoken dialog systems. Computer Speech and Language. Vol.21(2), 393–422.

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

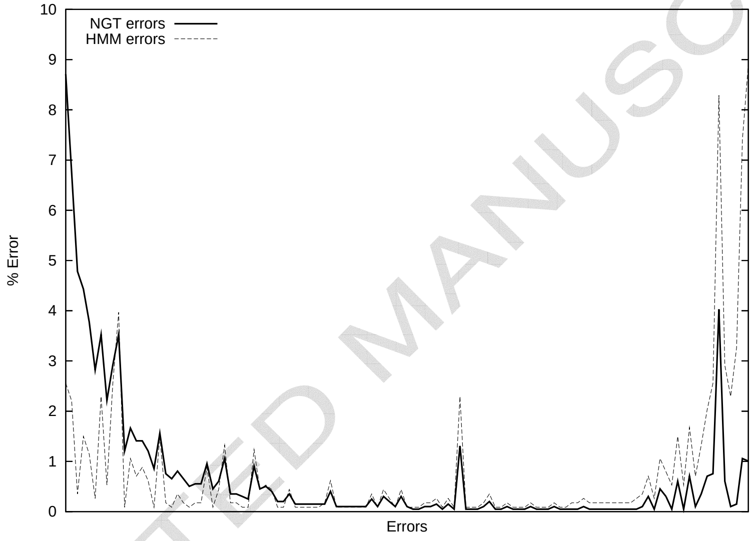


Figure 5: Error frequency for the user turn errors for the HMM-based model and the NGT model. The abscissa axis indicates the specific errors. The errors are ordered with respect to the difference of absolute occurrence in each model.

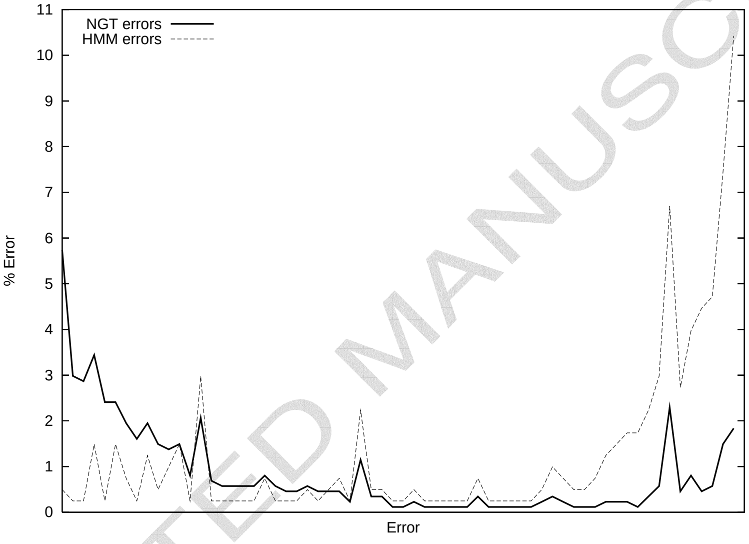


Figure 6: Error frequency for the user turn errors in the decoding task for the HMM-based model and the NGT model. The asbcissa axis indicates the specific errors. The errors are ordered with respect to the difference of absolute occurrence in each model.