Martínez-Hinarejos, C.; Benedí Ruiz, JM.; Tamarit Ballester, V. (2015). Unsegmented Dialogue Act Annotation and Decoding with N-Gram Transducers. IEEE/ACM Transactions on Audio, Speech and Language Processing. 23(1):198-211. doi:10.1109/TASLP.2014.2377595.

Unsegmented Dialogue Act Annotation and Decoding with N-Gram Transducers

Carlos-D. Martínez-Hinarejos, José-Miguel Benedí, *Member, IEEE,* Vicent Tamarit

***Abstract*—Most studies on dialogue corpora, as well as most dialogue systems, employ *dialogue acts* as the basic units for interpreting discourse structure, user input and system actions. The definition of the discourse structure and the dialogue strategy consequently require the tagging of dialogue corpora in terms of dialogue acts. The tagging problem presents two basic variants: a batch variant (annotation of whole dialogues, in order to define dialogue strategy or study discourse structure) and an on-line variant (decoding of the dialogue act sequence of a given turn, in order to interpret user intentions). In the two variants is unusual having the segmentation of each turn into the dialogue meaningful units (segments) to which a dialogue act is assigned. In this paper we present the use of the N-Gram Transducer technique for tagging dialogues, without needing to provide a prior segmentation, in these two different variants (dialogue annotation and turn decoding). Experiments were performed in two corpora of different nature and results show that N-Gram Transducer models are suitable for these tasks and provide good performance.**

***Index Terms*—Spoken dialogue systems, dialogue annotation, n-gram transducer**

I. Introduction

Dialogue systems are computer systems that interact with human users by means of dialogue, and are usually implemented to solve a given task. In the last decade, there have been many projects with the aim of developing a dialogue system, such like *Companions* [1], *Indigo* [2], *Classic* [3], or *PARLANCE* [4], among others. Most of the developed dialogue systems are devoted to a specific task, although some projects had as an initial aim developing dialogue systems for a variety of purposes. Speech is the usual input, but the multimodal paradigm is gaining attention in the last years [5], [6].

The core component of a dialogue system is the dialogue manager. It decides, by using the so-called dialogue strategy, how to manage the user input and which output must produce, also taking into account the previous development of the dialogue (dialogue history).

The definition of dialogue strategies can take benefits from the study of dialogue structure. This is an important research topic in the field of natural language processing. Although those studies usually focus in human-human dialogues analysis [7], [8], [9], some of their results could be used in the definition of dialogue systems.

All authors are with Universitat Politècnica de València, Pattern Recognition and Human Language Technologies Center, Camino de Vera, s/n, 46022, Valencia, Spain, {cmartine,jbenedi}@dsic.upv.es, tamarit@gmail.com

Dialogue structure is defined in the terms of the units that form a dialogue. The most evident unit in a dialogue is a *turn*, which can be defined as an interval of expression by a single dialogue participant. Inside a turn, one or more *segments* may appear. A dialogue segment (or *utterance* [10]) is any subsequence in a dialogue turn that has a relevant paper in the dialogue process.

There are many proposals for determining the dialogue structure or the dialogue management, but they can usually be divided into two classes by their approach to the resolution of this problem: rule-based approximations and data-based (or statistical) approximations. In the first case, there is a set of manually defined rules which determine the structure and govern the dialogue system. In the second case, a statistical model is in charge of deciding the dialogue structure and the system response. Both approximations have their advantages and drawbacks: the rule-based approach does not require a large amount of labelled data to define the dialogue strategy, but it is hard to define and adapt the set of rules for a specific task (it requires a human expert), whereas in the data-based approach the inference of the models parameters are directly derived from tagged dialogues, but requires a large set of tagged dialogues.

The concept of dialogue tagging that arises in the data-based approach consists of applying a set of labels that models the discourse structure during the dialogue. These labels allow to represent the current state of the dialogue and the possible actions that can be performed by a dialogue system at each point of the interaction. Therefore, it appears an association between a label and a communicative function or a dialogue situation, which allows the statistical models to become defined in terms of the set of labels. Since only the relevant information (from the dialogue viewpoint) is represented in the label (and consequently used by the model), the models become simpler. There have been many proposals to define the set of labels, but the most widely accepted proposal is tagging based on dialogue acts (DAs). A dialogue act is defined as a label that represents the meaning of a dialogue segment.

The problem addressed in this work relates to the tagging of dialogue units (whole dialogues or single turns) in terms of DAs. Two different scenarios appear:

- *Annotation*: the final aim is to infer the parameters of the models from a set of dialogues, which requires to tag the training data following the proposed set of DAs; thus, given a set of dialogues where segmentation is usually not available but the whole dialogue transcription is available, segments must be identified and DA labels must be properly assigned to each segment; manual tagging of

dialogues is hard and time-consuming, and consequently statistical models (inferred from a manually annotated subset of dialogues) can be used for obtaining a draft tagging of the dialogue corpus and reduce the global tagging effort.

- *Decoding*: the final aim is to encode all the interactions that appear during the dialogue system operation into DAs; in this case, segmentation is unavailable as well, and only the current and previous dialogue turns are available; the labels sequence must be determined, but their specific position is usually not necessary for the dialogue strategy; this is usually the case of a human-computer dialogue system in operation, where each user input and each system action must be interpreted in terms of DAs, and a model to decode user turns into DAs is necessary.

This article presents the use of a statistical technique, the N-Gram Transducer (NGT), in dialogue annotation and turn decoding for speech input. This technique is based on the GIATI technique [11] originally developed for Machine Translation. In contrast other techniques that use to employ generative models, NGT models the posterior probability. Previous preliminary results where obtained for the NGT technique for dialogue annotation [12], with competitive results with respect to other more classic techniques, but no turn decoding implementation has been described. Moreover, NGT can be directly used for obtaining the DA sequence without using a previous speech recognition (coupled decoding). In this work, experiments on NGT were performed in both the annotation and the decoding framework for two dialogue corpora of different nature (human-human and human-computer).

Section II summarises the different systems, annotation schemes, and statistical models that have been proposed for dialogue annotation and turn decoding. Section III presents the NGT technique. Section IV describes the data and Section V the assessment measures. Section VI presents the baseline system, whereas Section VII shows the experimental part for a human-human dialogue corpus, and Section VIII defines the experimental framework and results for a human-computer dialogue corpus. Section IX offers the conclusions and the future lines that can be explored for this technique.

II. Related Work

A. Dialogue models

Dialogue models employed in dialogue analysis (especially for human-human dialogues) and dialogue management are usually divided into rule-based models and data-based models. Rule-based models have been used in human-human dialogue assignment for obtaining local dialogue structure and dialogue acts [13] and for segmenting dialogues [14]. Rule-based models for dialogue management define the dialogue strategy by means of hand-crafted rules, which is the case of that presented in [15]. Some rule-based models employ hybrid approaches that use a set of tagged dialogues to identify the current state of the dialogue [16].

Data-based models have been used in human-human dialogue analysis in tasks like dialogue act assignment [10], style detection [17], or disfluency detection [18]. Most of these previous works usually assume the segmentation of the dialogue turns into dialogue segments. In dialogue management, data-based models use probabilistic models to define the dialogue strategy, i.e., given the dialogue history (including last user interaction) and, possibly, other environmental factors, they are given as inputs to a probabilistic model that provides as a result which is the next action to be performed by the system. In the last decade, the most used probabilistic models have been Markov Decision Processes (MDP) [19], [20] and Partially Observable MDP (POMDP) [21], [22], [23].

In this work we propose the use of the NGT model for tasks such as dialogue act assignment in unsegmented dialogue turns and dialogue act interpretation of user turns that can appear during the use of a dialogue system. Dialogues annotated with NGT models could be used in the estimation of the parameters of the probabilistic models (as MDP and POMDP), whereas DA interpretation of turns via NGT can be used as input of these models in the dialogue system operation.

B. Annotation schemes

Most of the models described in Subsection II-A rely on the interpretation of dialogue interactions (user input and system actions for human-computer dialogues) in terms of some dialogue-structure labels, which avoid redundant and unimportant information for the dialogue. As presented in Section I, the most usual approximation is tagging based on dialogue acts [24], but there have been many other proposals [25], [26].

A DA set can be used in human-human and in human-computer dialogues, and is usually tailored to the context of the human-human dialogues or to the task of the corresponding dialogue system. This fact caused the definition of different DA tagging schemes in the last twenty years. Some examples are DAMSL [27], VerbMobil [28], or DATE [29]. DAMSL is one of the most popular schemes, and it has been adapted for their use on the tagging of human-human corpora such as SwitchBoard [30] (using the SWBD-DAMSL variant [31]) or human-computer corpora such as AMITIÉS [32]. The DATE scheme has been applied as well to human-human dialogues and human-computer dialogues [33]. Therefore, adopting a set of DAs does not limit its application to an only type of dialogue corpus (human-human or human-computer).

Of course, different sets of DA tagging schemes can cause different effects in the DA detection (as results in Sections VII and VIII show), since estimating associations between dialogue situations and DA labels can be more difficult for some schemes (e.g., large number of DAs or ambiguous definition of DAs). In the last years, an effort on DA schemes standardisation has been performed, and an ISO proposal for a common set of DA labels and tagging rules has been developed [34].

Our proposal based on NGT is not related to the final use of the DA sequence, but to its correct detection in unsegmented dialogue turns. Thus, it could be applied to human-human dialogue studies and to human-computer dialogue systems, where segmentation is not usually available.

C. Automatic annotation and decoding

Whatever is the chosen DA scheme, two important tasks (as described in Section I) appear when developing data-based applications on dialogue: the *annotation* task (which is the aim of works such as [10], [35], [36]) and the *decoding* task.

Most previous works assumed that the start and end of the segments that compose the different turns are available, and therefore only DA classification for each segment is necessary. This segment classification can be done using different models.

In human-human dialogues, Stolcke *et al.* use in [10] a classic approach based on Hidden Markov Models (HMM) and n-grams on the SwitchBoard corpus [30], that is taken as baseline for many experiments. Apart from this seminal work, other proposals appeared in the latest years [37], [38], [39]. In human-computer dialogues there are several proposed models: Maximum Entropy [40], Spectral Clustering [41], Bayesian Networks [42], regression trees [43], SVM and Latent Semantic Analysis [44].

The assumption of the segmentation (i.e., knowing when the dialogue segments start and end inside a turn) is not realistic in some cases, such as for spoken dialogue systems and when only the transcriptions of the dialogue are available. In these situations, turns are usually distinguishable, but segments into the turn must be obtained with different techniques.

This segmentation task has been researched in several works [45], [46], [47], [48], [49]. However, not so many works consider together the segmentation and annotation of dialogue turns (i.e., the *unsegmented annotation problem*). Some authors propose a decoupled approximation [50], [51]. Only a few works, such as [52], [53], [54], [36], propose schemes in which segmentation and annotation is produced in a coupled process. These works use combinations of different models (A* search and n-grams [52], DBN with GMMs, and conditional probability tables [54]). In any case, to the best of our knowledge, systems that use a coupled segmentation and annotation require, for spoken dialogue turns, a previous recognition of the turn. After this recognition, the words can be used as input for the segmentation and annotation module.

In the case of our NGT proposal, only models of similar nature are employed, and the turn DA sequence can be directly obtained from speech.

III. The N-Gram Transducers Technique

The problem of obtaining the interpretation of dialogue turns (i.e., the corresponding sequence of DA labels) can be stated as an optimisation problem: given a word sequence $\mathcal{W}$ that represents a dialogue, obtain the sequence of DA labels $\mathcal{U}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$, that is:

$$\widehat{\mathcal{U}} = \underset{\mathcal{U}}{\operatorname{argmax}} \Pr(\mathcal{U}|\mathcal{W}) \tag{1}$$

A usual approximation is based on employing generative models (e.g., HMM). Formulation in this case is obtained by applying Bayes rule on Eq. (1), which decomposes the problem into:

$$\widehat{\mathcal{U}} = \underset{\mathcal{U}}{\operatorname{argmax}} \Pr(\mathcal{U},\mathcal{W}) = \underset{\mathcal{U}}{\operatorname{argmax}} \Pr(\mathcal{W}|\mathcal{U}) \Pr(\mathcal{U})$$

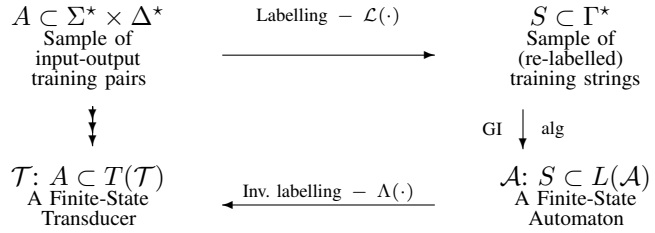


Fig. 1. General scheme for the GIATI technique. Σ, Δ and Γ are the input, output, and extended set of symbols, respectively. A and S are the initial sets of aligned and re-labelled samples. $L(\mathcal{A})$ and $T(\mathcal{T})$ represent the languages derived from $\mathcal{A}$ and $\mathcal{T}$, respectively. The GI algorithm is usually the inference of a smoothed n-gram, and $\mathcal{A}$ is the automaton equivalent to the inferred n-gram. $\mathcal{L}$ and Λ are the labelling and inverse labelling functions.

where $\Pr(\mathcal{W})$ is dropped because it does not depend on the maximisation variable $\mathcal{U}$.

In contrast, the optimisation proposed in Eq. (1) could be achieved by estimating the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$ using a discriminative model. In this work we propose to model $\Pr(\mathcal{U}|\mathcal{W})$ by using the N-Gram Transducers (NGT) technique.

A. Fundamentals of the N-Gram Transducers technique

The NGT technique is based on a Stochastic Finite-State Transducer (SFST) inference technique known as GIATI [11]. Although GIATI was initially defined as a Machine Translation technique, it can be easily adapted to the dialogue framework using as input language the words ($\mathcal{W}$) or other input information, and as output language the DA labels ($\mathcal{U}$). Figure 1 shows a scheme of the general GIATI process.

The last step (converting grammatical model into a transducer) is difficult when using smoothed n-grams. Therefore it is preferable to avoid this last step and use the smoothed n-gram directly as a transducer, which gives the name to the model: N-Gram Transducer (NGT). This can be applied for monotone alignments.

In the case of dialogues, the input language is the sequence of words of the dialogue $\mathcal{W}$, and the output language is the sequence of DA of the dialogue $\mathcal{U}$. Since words in a dialogue are organised in turns, it is usual to express the optimisation problem in terms of turns. Thus, given a dialogue with T turns, we express its associated word sequence as $\mathcal{W} = W_1^T = W_1 W_2 \cdots W_T$ and the corresponding DA sequence as $\mathcal{U} = U_1^T = U_1 U_2 \cdots U_T$. Notice that W_t and U_t represent the sequences of words and DA, respectively, for the turn t. Consequently, without losing generality, the posterior probability could be stated as:

$$\Pr(\mathcal{U}|\mathcal{W}) = \Pr(U_1^T|W_1^T) \approx \prod_{t=1}^{T} \Pr(U_t|W_t) \tag{2}$$

where we assume that the sequence of labels of one turn depends only on the words of that turn.

Alignments are local for each turn (i.e., words of W_t only can be aligned to DA of U_t), and define the *segments* (sequence of consecutive words assigned to a DA label) of the turn. Given a turn $W_t = w_1^l = w_1 w_2 \ldots w_l$, a segmentation on

Yes , uh , I don't work , though , but I used to .		
↑	↑	↑
%	sd	sd
Yes , uh ,@% I don't work , though ,@sd but I used to .@sd		

Fig. 2. An alignment between a dialogue turn and its corresponding DA labels (from the SWBD-DAMSL scheme, %: uninterpretable, sd: statement-non-opinion), and the result of the re-labelling process, where @ is the attaching metasymbol.

W_t is defined by a sequence of indexes $(s_0, s_1, s_2, \ldots, s_r)$, where r is the number of segments, $s_0 = 0$, $s_r = l$ and $W_t = w_{s_0+1}^{s_1} w_{s_1+1}^{s_2} \ldots w_{s_{r-1}+1}^{s_r}$. If $U_t = u_1 u_2 \ldots u_r$, each segment $w_{s_{i-1}+1}^{s_i}$ is assigned to the DA u_i. Alignments are established between the DA label and the final word of its segment.

By using this segmentation notation and focusing in a certain turn ($U = U_t$ and $W = W_t$), the calculation of the terms $\Pr(U|W)$ of Eq. (2) can be approximated as:

$$\Pr(U|W) \approx \sum_{r,s} \prod_{i=1}^{r} \Pr(u_i | w_{s_{i-1}+1}^{s_i}) \tag{3}$$

where all possible number of segments r and segmentation s are considered and, for a given segmentation s of r segments, it is assumed that the probability of assigning u_i to the i-th segment only depends on the words of that segment ($w_{s_{i-1}+1}^{s_i}$).

Given the sequence of words of a turn t, $W_t = w_1 w_2 \ldots w_l$, and the corresponding DA sequence $U_t = u_1 u_2 \ldots u_r$, the labelling function that is usually used attaches the DA label u_i to the last word of the i-th segment in W_t (w_{s_i}) using a metasymbol (e.g., @).

Consequently, the terms in the product of Eq. (3) can be expressed as:

$$\Pr(u_i | w_{s_{i-1}+1} \ldots w_{s_i}) \approx \mathrm{q}(w_{s_{i-1}+1} \ldots w_{s_i} @ u_i) \tag{4}$$

where q is a score function on the sequence that does not consider the normalisation term of the probability. The posterior probability could be estimated in an n-gram fashion, given as final estimator:

$$\Pr(u_i | w_{s_{i-1}+1} \ldots w_{s_i}) \approx \mathrm{q}(w_{s_i} @ u_i | w_{s_i-n}^{s_i-1}) \prod_{k=s_{i-1}+1}^{s_i-1} \mathrm{q}(w_k | w_{k-n}^{k-1}) \tag{5}$$

Notice that in the terms $\mathrm{q}(w_k | w_{k-n}^{k-1})$, the sequence of words could be extended to previous segments in order to have a proper definition of the score.

The use of the labelling function that attaches the DA label to the last word of each segment produces as a result, for turn t, a extended word sequence $e_1 e_2 \ldots e_l$, where:

- $e_i = w_i$ when w_i is not aligned to any DA.
- $e_i = w_i @ u_k$ when w_i is aligned to the DA u_k.

Figure 2 presents an example of alignment for a dialogue turn and the corresponding extended word sequence. After this step, the n-gram can be inferred from the total set of extended word sequences to build the NGT model. Notice that terms in Eq.(5) relate to the NGT model by the association of the observation score $\mathrm{q}(w_{s_i} @ u_i | w_{s_i-n}^{s_i-1})$ to the sequences of the end of the i-th segment and the association of the transition scores (those of the product) to the previous sequences in that segment.

The search process (decoding) in the NGT model is the key point of the technique. The decoding is applied to a sequence of input words (without DA labels) and provides the sequence of the DA labels and their positions with respect to the input sequence. The decoding is a search process in the NGT model, whose search space is modelled as a tree where each node has associated a score. The i-th level of the tree corresponds to the i-th word in the input, and each input word is expanded for all the possible outputs it has associated in the alignments in the training corpus. For example, if the input word w_i was aligned in the training corpora to outputs (DA) d_1, d_2, and d_3, apart from the empty output, all the nodes in level $i-1$ will expand into four children nodes, each of them associated to the corresponding output (i.e., w_i, $w_i @ d_1$, $w_i @ d_2$, $w_i @ d_3$).

Apart from that, in the NGT decoding an n-gram of DA labels is used to compute the score of the nodes of words with associated outputs. Thus, the score of a node is basically calculated as a product of three different factors:

1) The score of its parent node p_P.
2) The score of the sequence of the n extended words that finish in the node (which is given by the NGT model, an n-gram of degree n), $\mathrm{q}(e_{i-(n-1)}, \ldots, e_i)$.
3) If the node corresponds to an extended word with associated output (i.e., $e_i = w_i @ d_k$), the score of the sequence of the m DA that finish in the node (given by an n-gram of degree m of DA sequences, inferred from the training data), $\mathrm{q}(d_{k-(m-1)}, \ldots, d_k)$.

Therefore, for a expanded node with empty output, its associated score is calculated by $p = p_P \cdot \mathrm{q}(e_{i-(n-1)}, \ldots, e_i)$, and for the rest of the nodes is calculated by $p = p_P \cdot \mathrm{q}(e_{i-(n-1)}, \ldots, e_i) \cdot \mathrm{q}(d_{k-(m-1)}, \ldots, d_k)$. In Figure 3 an example of tree search is provided.

There are a few similarities between NGT and the Hidden Event Language Model (HELM) proposed in [18], [55], [56], such as the use of n-gram and the appearance of events that are somewhat hidden in the processed sequence. However, while in HELM the hidden events appear at the same level than the visible events and are removed in the estimation of the n-gram probabilities, in NGT the output labels are added to the input tokens and the extended symbol forms part of the probability estimation. Moreover, HELM does not explicitly models the relations between hidden events in the sequence, while NGT decoding does it by using the n-gram of output labels.

The basic NGT search can use alternatives such as:

- Beam-search to avoid the exponential growth of the search space.
- Limited expansion of nodes, by expanding only the k children nodes of highest score for each parent node.
- Weight factors that promote or penalise the nodes with output or that alter the influence of the NGT model with respect to the n-gram of DA; in this last case, the Output Grammar Scale Factor (OGSF) is defined as a factor that increases or reduces the contribution of the n-gram of DA to the score calculation for each node.

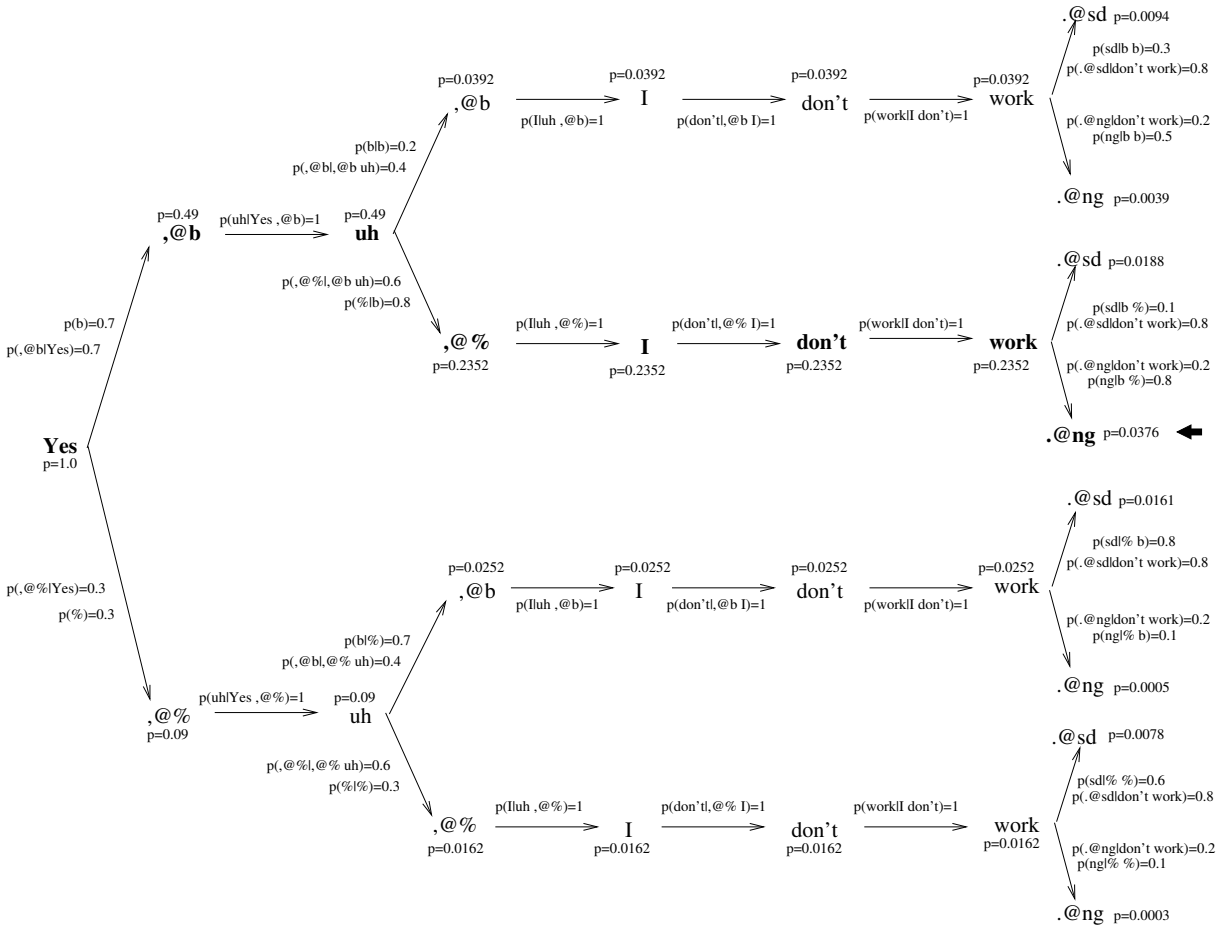


Fig. 3. An example of the tree search for the NGT model. In this example, both the NGT and the n-gram of DA are modelled by 3-grams. In the nodes where an output is produced, its score is computed from the score of the parent node, the NGT probability and the n-gram of DA probability. Best hypothesis is in boldface and marked by a dark arrow. The meaning of the DA labels is uninterpretable (%), backchannel (b), statement-non-opinion (sd) and negative-non-no-answers (ng).

B. N-Gram Transducers on dialogue

The NGT technique can be applied to several tasks in dialogue. In this work we focus in the tasks described in Subsection II-C: dialogue annotation, and decoupled/coupled turn DA decoding.

- Dialogue annotation: in this case, the input of the model is a transcription of a complete dialogue, which can include punctuation marks and annotation of non-linguistic phenomena (such as laughter, coughing, etc.). In this task, the separation between turns is commonly available, but the segmentation into segments is not usual. Since segments are usually confined in a turn, the last word of a turn must be the last word of a segment. Consequently, the final word of a turn will have assigned a DA. The tree expansion is done for the complete sequence of words. The final solution is the optimal for the whole dialogue, and provides the segmentation and annotation with DA of all the turns of the input dialogue. Therefore, this option can be used in the annotation of dialogue corpora.
- Decoupled turn decoding: in this case, the input of the model is a recognition result of a dialogue turn; in this case it is unusual to have some special tokens (such as punctuation marks) in the input. In this task, the dialogue turns previous to the current input are known and their decoding is fixed. Therefore, those previous turns (dialogue history) are modelled as a single branch in the search tree and when the new input is processed children nodes for only that branch are created. The final solution is a single branch which is optimal for the turn but without knowing the future turns. From this branch the turn segmentation and annotation is obtained, although for the decoding process only the sequence of DA is needed. Therefore, this option can be used to decode user intentions from the result of a speech recogniser.
- Coupled turn decoding: this case is similar to the previous, but consists of using directly the NGT model as language model in the speech recogniser (i.e., coupled speech recognition and DA decoding). Besides, the speech recogniser must take into account the n-gram of DA in the recognition process; therefore, the speech recogniser must be slightly modified to take into account this n-gram when calculating the probabilities for the NGT model states with associated output. This option can be used to directly obtain DA labels in speech-

TABLE I
SUMMARY OF THE FEATURES OF THE SWITCHBOARD AND THE DIHANA CORPUS. FOR DIHANA, U MEANS USER TURNS AND S SYSTEM TURNS.

Corpus	SwitchBoard	Dihana
Language	English	Spanish
Nature	Human-human	Human-computer
Semantically restricted	No	Yes
Number of dialogues	1155	900
Number of turns	115,000	6280 U + 9133 S
Vocabulary size	42,000	900
Annotation scheme	SWBD-DAMSL	IF-DIHANA
Number of DA labels	42	248

Spk	Seg	Transcription	Lab
S1	S1-1	Yeah,	aa
	S1-2	to get references and that,	sd
	S1-3	so, but, uh,	%
	S1-4	I don't feel comfortable about leaving my kids in a big day care centre, simply because there's so many kids and so many <sniffing> <throat_clearing>	sd
S2	S2-1	I think she has problems with that, too.	sd

Fig. 4. An example of annotated turns in the SwitchBoard corpus. The meaning of the labels is statement-non-opinion (sd), uninterpretable (%) and agree/accept (aa).

input dialogue systems, which is supposed to be more accurate and efficient than the decoupled option (which was previously described as decoupled turn decoding).

In this work, we will centre in these previously described applications, although NGT can be used for other applications, such as obtaining a draft segmentation that can be used by other DA assignment models [57].

IV. DIALOGUE DATA

In this section we present the experimental data that we used in the different experiments with the NGT technique. Two different corpora are introduced: the SwitchBoard corpus [30] and the Dihana corpus [58]. These two corpora present dissimilar features with respect to the size of the vocabulary, the set of DA labels, the nature of the interaction, the associated semantic restrictions, etc. A summary of the different features of these two corpora is presented in Table I.

A. SwitchBoard corpus

The SwitchBoard corpus [30] is a popular corpus of telephone conversations in English between two human speakers. The speakers discuss about general defined topics, but without a clear task to accomplish (they can discuss about politics, economics, social matters, etc., but without a defined objective). The corpus recorded spontaneous speech, with frequent overlaps and interruptions between the speakers, and with several spontaneous speech phenomena (such as hesitations, non-linguistic sounds, etc.) and background noises.

The corpus consists of 1155 conversations (approximately 115,000 turns), that were manually transcribed with a special notation for spontaneous speech phenomena (specially overlaps and non-linguistic sounds). Vocabulary size is about 42,000 words. Total recorded speech signal is about 95 hours.

The set of DA labels that were used in the annotation of the Switchboard dialogues is known as SWBD-DAMSL [31]. This set is a simplified version of the standard DAMSL annotation set [27]. SWBD-DAMSL comprises a total number of 42 different labels, which cover communicative functions such as statement, question, backchannels, etc., but with associated subtypes (e.g., statement-opinion and statement-non-opinion). In the annotation process, each turn was segmented into the corresponding segment (an average of 1.7 segments per turn was obtained) and a label was associated to each segment according to a set of rules. The manual labelling was performed by 8 different human experts, given an inter-annotator agreement with a Kappa value of 0.8 [10]. An example of annotation is provided in Figure 4.

Even though its intrinsic difficulties, SwitchBoard has become in the last years a sort of standard corpus to evaluate annotation tools and models. Several works used SwitchBoard to evaluate their proposed models, such as [10], [59], [36]. Some of these previous works do not give details on the corpus preprocessing or experimental framework.

In our case, the preprocessing of the SwitchBoard corpus removed interruptions and overlaps by joining the separated pieces of turns (consequently, all the segments in the turn remain in a single speaker interaction), all the punctuation marks were separated as single words, and all the words were transcribed to lowercase[1]. This preprocessed version can be used in the dialogue annotation experiments. We built another version without punctuation marks at all, in order to simulate the output of an hypothetically "perfect" speech recogniser and to use it in turn decoding experiments.

B. Dihana corpus

The Dihana corpus [58] is a set of 900 dialogues in Spanish. The dialogues were acquired in a Wizard-of-Oz (WoZ) environment [60] which simulates a human-computer interaction. These dialogues were directed by the definition of scenarios in a task related to railway information (timetables and fares) for long-distance trains in Spain. The acquisition was only restricted with respect to the objectives of the scenarios, and no other syntactical, lexical or semantic restriction was applied in the interaction with the users.

The corpus acquired conversations from 225 speakers (153 male and 72 female) that presented small Spanish dialectal variants. The total number of turns in the corpus is 6280 for the users and 9133 for the system, with a vocabulary of approximately 900 words. The total amount of acquired signal is about 5.5 hours. All the dialogues were manually transcribed, including special annotation marks for spontaneous speech phenomena. No overlaps or interruptions were present in the corpus, since in the acquisition process the user was not allowed to interrupt the system prompts (possible interruptions were ignored). The spontaneous speech marks were removed for obtaining the final version of the corpus to be annotated.

[1]Preprocessed corpus available at `www.dsic.upv.es/~cmartine/research/resources.html`.

Spk	Seg	Transcription		
		Lev 1	Lev 2	Lev 3
S	S1	Welcome to the railway information system. How may I help you?		
		Open	Nil	Nil
U	U1	I want to know the departure times from DEP-TW(Valencia)		
		Que	Dep-h	Org
	U2	to Madrid		
		Que	Dep-h	Dest
	U3	arriving on DATE(May the 15th of 2004).		
		Que	Dep-h	Day
S	S2	Do you want to leave on DATE(Sat, May the 15th of 2004)?		
		Conf	Day	Day
U	U4	Yes.		
		Accept	Day	Nil
S	S3	Consulting times for trains from DEP-TW(Valencia) to DST-TW(Madrid) on DATE(Sat, May 15th 2004).		
		Conf	Dep-h	Dest, Day, Org
	S4	Wait a moment, please.		
		Wait	Nil	Nil
	S5	There are TR-NUM(several) trains. The TR-ORD(first one) leaves at DEP-HR(7:45) and arrives at ARR-HR(11:14), and the TR-ORD(last one) leaves at DEP-HR(18:45) and arrives at ARR-HR(22:18).		
		Ans	Dep-h	Arr-h, Dep-h, Ord-n, N-tr
	S6	Do you need anything else?		
		Cons	Nil	Nil

Fig. 5. An excerpt of an annotated dialogue (translated from Spanish into English) from the Dihana corpus. *Nil* denotes the absence of information. Words in capital letters denote categories.

The set of DA labels that was defined for the annotation of Dihana is an adaptation of the Interchange Format (IF) used for dialogue annotation [61]. The IF format defines three different levels for each label, called respectively speech act, concept, and argument. This set was adapted for Dihana [62], giving a total of 248 different labels (152 for user turns and 95 for system turns). Figure 5 presents an annotation example.

In contrast with SwitchBoard, Dihana is not such a sort of standard dialogue corpus, but it presents several interesting features for the study of the dialogue process and the implementation of actual dialogue systems: it is a medium-size corpus, task-oriented, with vocabulary limited to the task. In conclusion, Dihana is a useful corpus to complement the conclusions obtained with SwitchBoard and check the models in a real dialogue system framework

The preprocessing of the Dihana corpus consisted in lowercase transcription, separation of punctuation marks, adding a speaker mark (U for user and S for system) to each word and a categorisation of the most frequent semantic categories that are present in the task (such as town names, hours, dates, etc.)[2]. This preprocessed corpus can be used in the dialogue annotation experiments. Apart from this, a version without punctuation marks and the output of a speech recogniser are available for decoupled turn decoding. Audio data is available and allows coupled turn decoding experiments.

[2]Preprocessed corpus available at `www.dsic.upv.es/~cmartine/research/resources.html`.

V. Assessment measures

The assessment of the results depends on the specific task to be performed. When using the models for the annotation of whole dialogues, it is important to have the correct labels in the correct position of the turns. However, for turn decoding, only the correct sequence of DA labels is important, since the label position in the turn is not usually provided to the dialogue manager.

Previous works defined several assessment measures [63]. Apart from these, we propose alternative measures based on edit distance. Each of these measures has a specific purpose (i.e., measure the quality of different things). According to this purpose, they can be divided into three different groups:

1) *Decoding measures*: only DA labels, no positions
   - CER (Classification Error Rate): it is a classical measure which is used when each unit to be annotated is a segment (i.e., the segmented case), and only one DA label is assigned; it measures the percent of errors committed in the classification.
   - DAER (DA Error Rate): it is our proposed measure; it considers whole turns, which may have several DA labels (i.e., the unsegmented case); therefore, the sequence of DA labels from the reference is compared with the sequence of DA labels obtained from the decoding process; DAER measures the edit distance between these sequences; DAER overcomes the limitation of the CER measure on the segmented case.
2) *Segmentation measures*: only positions of the DA labels in the turn are compared:
   - NIST-SU: it is classical proposal which computes number of segmentation errors (missed segments and false alarm segments) divided by the number of segments of the reference; its limitation is that does not consider position substitutions.
   - DSER (DA Segmentation Error Rate): it is a classical measure that computes the number of segments of the reference incorrectly segmented divided by the total of segments of the reference; the main difference with NIST-SU is that it only takes into account reference segments, ignoring errors produced by an excessive segmentation in system output; its limitation is that takes segment as whole sequence, not as limits.
   - SegER (Segmentation Error Rate): it is our proposed measure for segmentation error; it is computed as the edit distance between sequences of reference positions and annotation positions (those obtained by the system); SegER is proposed to avoid the NIST-SU and DSER limitations, since accounts for the fact that in real annotation a wrong segmentation can be corrected by using only final boundaries (initial boundaries can be supposed to be the next position to the previous boundary), by inserting, deleting or moving the corresponding boundaries (which correspond to edit operations).

Decoding measures (label)					
Reference	B	Z	K	B	Q
System seg.	B	Z	Z	B	Q
System		Z	Z	B	Q
CER	✓	✓	×	✓	✓
DAER	D	C	S_1	C	C

Segmentation measures (position)											
Reference	B	Z	Z	Z	K	K	K	B	Q	Q	|
System	Z	Z	Z	Z	Z	Z	B	B	Q	Q	|
NIST-SU	✓			×		×	×	✓		✓	
DSER	✓			×			×	×		✓	
SegER	C			D		S_1	S_1	C		C	
Annotation measures (label and position)											
Lenient	×	✓	✓	✓	×	×	×	✓	✓	✓	
Strict	×	×	×	×	×	×	×	×	✓	✓	
SegDAER	S_1			D		S_2	S_2	C		C	

Measure	Error computation
CER	1 Err/5 Ref = 20%
DAER	(1D+1S)/(3C+1D+1S)=40%
NIST-SU	3 Err/5 Ref = 60%
DSER	3 Err/5 Ref = 60%
SegER	(1D+1S)/(3C+1D+1S)=40%
Lenient	4 Err/10 Ref = 40%
Strict	8 Err/10 Ref = 80%
SegDAER	(1D+2S)/(2C+1D+2S)=60%

Fig. 6. An example of how to calculate the different assessment measures. Reference and system show the DA labels present in the reference and given by the system (B, Z, K, and Q represent DA labels, | represents segment limits). In NIST-SU, DSER, *Lenient*, *Strict*, and CER measures, × means error and ✓ correct. In SegER, SegDAER, and DAER, S_k means substitution (two S_k with the same k value represent the same substitution), D deletion, and C correct.

3) *Segmentation and annotation measures*: both DA labels and their positions are considered:
   - *Lenient*: it is a classical measure that calculates the number of words with incorrect DA label divided by the total number of words.
   - *Strict*: it is a classical measure that calculates the number of words with incorrect DA label or incorrect segmentation divided by the total number of words; the difference with *Lenient* is that takes into account if the word is in the correct segment.
   - SegDAER (Segmentation and DA Error Rate): it is our proposed measure, in which the units to be compared are composed of the position joined to the DA label; again, reference sequence is compared against annotation sequence with the edit distance to obtain its value; SegDAER is proposed to avoid a feature common to *Lenient* and *Strict*, which is that they consider all the words of the segment affected by the label and the boundaries; however, in an annotation framework, labels can be considered to be at the end of the segment (and they would be interpreted to affect all the words of the segment), and correcting them (in position or in value) would imply edit operations on the label, not on the total sequence of words.

Figure 6 (very similar to that used in [63], [50], [64]) shows how sequences are obtained and compared for each of the proposed measures. In the annotation experiments, errors can be produced by an incorrect position (NIST-SU, DSER, and SegER give a measure on that) or by an incorrect label (DAER gives a measure on this type of error); *Lenient*, *Strict*, and SegDAER are measures that take into account these two sources of errors altogether, and are appropriate to evaluate the performance in the annotation experiments. However, in decoding evaluation position errors are not important, thus making the DAER measure the most appropriate to evaluate the quality of the decoding process.

VI. Baseline system

In this section we present a set of baseline models: a generative model based on Hidden Markov Models (HMM) with n-grams as language model of DA (model similar to that proposed in [10]), a discriminative model based on Conditional Random Fields (CRF) [65], such as that used in [66], and a sequential model (i.e., apply a model for segmentation and then one for DA assignment on the obtained segments, which is an usual option [71]) that uses CRF for segmentation and Support Vector Machines (SVM) for DA assignment. The performance of the NGT model for the annotation in the unsegmented case will be compared against these models. We present a comparison of the HMM-based model with other authors work in Subsection VI-A, the experimental framework in Subsection VI-B, and the results in Subsection VI-C.

A. Comparison with previous work

In order to validate our implementation of the HMM baseline model, we compared other authors results with the results given by this model. Although we are interested in the unsegmented case, most authors present only classification of segments (i.e., segments are given and only a single DA label must be assigned to the segment). Thus, for this comparison we used a HMM-based model version that acts on segmented input (i.e., takes the whole segment and classifies it into any of the different classes represented by the HMM models).

This comparison is made on the SwitchBoard corpus, which has been extensively used in DA tagging since its publication. Many works, such as [10], [35], [67], [68], [69], [9], [70], have used this corpus. Unfortunately, each of them uses a different experimental framework (most cases not fully specified), and assume the segmentation of the corpus (in contrast with the unsegmented case, which is our main interest).

Anyway, results of different authors are presented in Table II, along with our baseline results[3]. The 11-fold cross-validation partition was chosen in order to have the same number of dialogues (105) in each partition.

Results show that our HMM proposal obtains a performance similar to that obtained by other authors, although the different composition of test and training sets makes the different results not directly comparable (except in the case of Stolcke partitions). Therefore, we will take this implementation as baseline system for evaluating the quality of the NGT results.

B. Experimental framework

Our experimental framework is based on using a cross-validation approach. In the case of the SwitchBoard corpus,

[3]See `www.stanford.edu/~jurafsky/ws97/` for Stolcke partitions and `www.dsic.upv.es/~cmartine/research/resources.html` for 11-fold partition.

TABLE II
COMPARISON OF CER RESULTS FOR THE SWITCHBOARD CORPUS FOR DIFFERENT MODELS AND TEST FRAMEWORKS. LAST TWO RESULTS REFER TO OUR BASELINE HMM-BASED MODEL.

Author	Error	Conditions
[10]	29.0%	Stolcke partitions
[59]	28.7%	Unknown 4k test
[67]	29.7%	Unknown 10-fold cross validation
[68]	30.2%	Unknown test, ICSI training
[9]	30.0%	Unknown test, 49 labels
[70]	29.6%	Unknown test, 173 dialogues
HMM-baseline	30.3%	Stolcke partitions, 3-gram
HMM-baseline	29.5%	11-fold cross validation, 3-gram

TABLE III
Lenient, *Strict*, AND SEGDAER SWITCHBOARD AND DIHANA RESULTS WITH HMM AND DIFFERENT DA n-GRAMS, WITH CRF, AND WITH CRF SEGMENTATION AND SVM TAGGING. CONFIDENCE INTERVALS LOWER THAN ± 0.2 IN SWITCHBOARD AND THAN ± 0.6 IN DIHANA.

		HMM / DA n-gram				CRF	CRF+ SVM
		2	3	4	5		
SWBD	*Lenient*	30.9	31.0	31.2	31.6	38.5	**28.9**
	Strict	79.7	79.5	79.8	80.3	63.7	**54.0**
	SegDAER	60.7	60.5	61.1	62.0	48.9	**41.9**
Dihana	*Lenient*	13.2	**13.1**	13.4	13.7	24.4	13.8
	Strict	61.4	61.5	61.7	61.9	25.0	**14.6**
	SegDAER	33.7	34.5	34.6	35.2	26.5	**22.2**

we used 11 partitions of 105 dialogues each partition. In the case of the Dihana corpus, since it presents different features, partitions are 5 of 180 dialogues each partition[4].

The HMM experiments tested different values for the degree of the DA n-gram (from 2 to 5). The CRF experiments were performed by using CRF++[5] with CRF-L2 algorithm, $C = 1$, and $\eta = 5 \cdot 10^{-3}$; features were the word and whether it is final or not; the template file was that given with CRF++ for the Base-NP task. The sequential approach experiments used CRF segmentation (same parameters) and SVM tagging (with `libsvm`[6], cost 1, linear kernel).

To check the statistical significance, 90% confidence intervals were calculated using bootstrapping with 10,000 repetitions for all the experiments [72].

C. Baseline results

The baseline experiments are annotation experiments, i.e., the input are whole dialogues. Therefore, the search is performed from the first word in the first turn of the dialogue up to the last word of the last turn of the dialogue. Turn boundaries are taken into account since we consider that DA labels do not span between different turns. No other information on segmentation is given at the input of the baseline system.

Experiments were performed in the conditions described in Subsection VI-B. They produced the results presented in Table III. The sequential approach gives better results in all cases (except for Dihana *Lenient*). Examining in detail the results, this good behaviour is mainly caused by the more accurate segmentation given by the CRF model used only for segmentation. Moreover, the SVM model for DA assignment produces much better results on SwitchBoard because the lower number of classes (DA labels), which makes easier the association between word segments and DA label. In Dihana, where the number of labels is quite higher, improvements is not so high, although the input vocabulary is more reduced.

Unfortunately, as far as we know, the annotation experiment has not been performed by other authors in the SwitchBoard corpus (although many works used other corpora such as ICSI-MRDA [64], [68], [73], [50], AT&T VoiceTone [74], and SpeechDate [17]), and no clear comparison is possible with previous works in this unsegmented case.

[4]Partitions available at `www.dsic.upv.es/~cmartine/research/resources.html`

[5]`http://code.google.com/p/crfpp/`

[6]`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

VII. RESULTS IN HUMAN-HUMAN DIALOGUES

In this section we present the experiments directed to evaluate the NGT technique for dialogue annotation and decoupled turn decoding in human-human dialogues. Although our main interest is on human-computer dialogues and turn decoding, the use of human-human dialogues and annotation allows us to evaluate the performance of the NGT model with respect to the baseline systems presented in Section VI.

A. Experimental framework

Our experimental framework is that defined in Subsection VI-B. The complete experiments tested different values for the degree of the NGT n-gram and the DA n-gram (from 2 to 5). The Output Grammar Scale Factor parameter was optimised in the cross-validation.

B. Dialogue annotation experiments

In the dialogue annotation experiments, the conditions are the same that those described for the baseline system (Subsection VI-C) in terms of input type and segmentation information. Table IV shows the annotation task results. The best option is using an NGT model of degree 3 (significantly better than other NGT models), and no significant differences can be observed for the different DA language models employed.

Results show that SegDAER is a coherent measure with respect to *Lenient* and *Strict*, and that the SegER measure is coherent with NIST-SU and DSER. Moreover, it confirms that SegDAER results are not as optimistic as *Lenient* and not as pessimistic as *Strict*, giving a more balanced measure of the correction effort for annotation.

We can see that half of the annotation decisions are correct (SegDAER results, 49.2 for NGT 3, DA n-gram 3 and 4), which is significantly better than the HMM results in Table III, with an absolute difference of about 15% in *Strict* and 11% in SegDAER (*Lenient* results are worse in a 9%). When comparing with the CRF results (Table III), SegDAER differences are not significant, although CRF *Strict* and *Lenient* results are slightly better.

DAER, NIST-SU, DSER, and SegER results show that the main source of errors are incorrect labels, which is consonant

TABLE IV
SWITCHBOARD ANNOTATION RESULTS. FIRST COLUMN REFERS TO THE DEGREE OF THE DA n-GRAM AND SECOND ROW TO THE DEGREE OF THE NGT MODEL. CONFIDENCE INTERVALS ARE LOWER THAN ± 0.3 FOR ALL MEASURES AND EXPERIMENTS.

		NIST-SU				DSER				SegER			
	DA n-gram	2	3	4	5	2	3	4	5	2	3	4	5
NGT	2	**23.0**	**23.0**	**23.0**	**23.0**	**33.3**	33.4	33.4	33.4	**21.3**	**21.3**	**21.3**	**21.3**
	3	23.8	23.7	23.7	23.8	33.7	33.6	33.6	33.7	21.8	21.8	21.7	21.8
	4	26.0	26.0	26.0	26.1	35.6	35.6	35.6	35.7	23.3	23.3	23.3	23.4
	5	28.5	28.7	28.7	28.8	37.9	37.9	38.0	38.1	25.2	25.2	25.3	25.3

		Lenient				*Strict*				SegDAER				DAER			
	DA n-gram	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
NGT	2	41.0	50.6	50.7	50.7	66.2	66.1	66.1	66.1	50.7	50.6	50.7	50.7	47.3	47.2	47.2	47.3
	3	39.9	**39.5**	**39.5**	39.6	65.9	**65.7**	65.8	65.9	49.3	**49.2**	**49.2**	49.4	45.4	**45.2**	**45.2**	45.5
	4	41.5	41.0	40.9	41.1	68.6	68.4	68.4	68.5	52.3	52.1	52.1	52.4	47.9	47.6	47.6	47.9
	5	42.8	42.3	42.1	42.3	71.3	71.1	71.2	71.2	55.7	55.5	55.5	55.7	50.6	50.4	50.4	50.5

with the difficulty of labelling the corpus even for human users (as was pointed out in [10]). However, segmentation errors are higher than those provided by the sequential approach. Apart from that, since NGT is based mainly on last words of the segment, it does not has as many informational sources as the SVM model for determining such ambiguous labels. This explains why the sequential approach is not improved by NGT.

C. Decoupled turn decoding experiments

In these experiments we used two different versions of the corpus: with and without punctuation marks. The version with punctuation marks simulates a typed system or a speech recognition system which can produce perfect transcriptions. The version without punctuation marks simulates a perfect speech recognition system that cannot provide punctuation marks (that can be critical in the obtainment of the appropriate sequence of DA). The NGT model for the punctuation marks corpus is the same than for annotation, while in the other version the model is trained from dialogues without punctuation marks.

Table V shows the results for both versions of the SwitchBoard corpus. The differences between using the dialogue-by-dialogue approximation (used in annotation) and the turn-by-turn approximation (used in decoding) with punctuation marks are not significant (45.2 and 45.5 are the best results in the two cases, see Tables IV and V). Consequently, the turn-by-turn approximation can be used in the annotation experiments with similar results but with lower spatial complexity (since only one branch per turn is kept after the turn decoding).

However, when punctuation marks are not present, results degrade dramatically (from 45.5 to 51.4). This is explained by the nature of the NGT model, where the most important parameters are the last n words of the segment; therefore, deletion of punctuation marks (which are very usual at the end of segments) may have a large impact in the correct estimation of the NGT model parameters. Thus, we conclude that the presence of punctuation marks is critical to obtain better results for this type of dialogues (non-task-oriented, large vocabulary), and using this model with the direct output of a speech recogniser can be error prone. Results show that different DA language models have similar effect, with 3-grams as the best option for the NGT model (although there are no significant differences when using other NGT models).

TABLE V
SWITCHBOARD DECODING RESULTS (DAER). CONFIDENCE INTERVALS ARE IN ALL EXPERIMENTS LOWER THAN ± 0.3.

		With punctuation marks				Without punctuation marks			
	DA n-gram	2	3	4	5	2	3	4	5
NGT	2	48.2	48.1	48.1	48.1	**51.4**	**51.4**	**51.4**	**51.4**
	3	45.7	**45.5**	45.6	45.7	51.9	51.7	51.7	51.7
	4	47.0	46.9	46.9	47.0	52.4	52.4	52.4	52.4
	5	49.3	49.2	49.2	49.4	52.9	52.8	52.8	52.9

VIII. RESULTS IN HUMAN-COMPUTER DIALOGUES

In this section we present the experiments directed to evaluate the NGT technique for a human-computer system. We present the experimental framework for the Dihana corpus and the results for annotation and decoding.

A. Experimental framework

Our experimental framework for Dihana used the cross-validation approach described in Subsection VI-B. The partitions were applied in the annotation and decoding processes.

In the case of the decoding processes (decoupled and coupled), the user dialogue turns must suffer a speech recognition process. Acoustic and language modelling was performed on the 4 training partitions that correspond to each test partition (i.e., a set of 5 acoustic and language models were used). In speech experiments only user turns are taken into account (since there is no speech signal associated to system turns).

For Dihana speech recognition, feature vectors consists of 12 cepstrum values plus energy, and delta and acceleration coefficients. Acoustic modelling is performed by HMM with a three-state, left-to-right without skips topology, that model contextual units (triphones); a total set of 802 different triphones were employed for each set, with a total number of 19,182 gaussians for all the models for each partition. These models were trained by using the HTK toolkit [75].

With respect to language models, the inclusion of the dialogue context in its estimation and usage helps to obtain better results. Context refers to the presence of system turns that can be introduced in the search process to restrict the speech hypothesis. Additionally, the finite nature of each turn can be

used to disregard other hypothesis. This context inclusion is made differently in the decoupled and coupled framework:

- Decoupled: context is included by adding (for each turn) the previous system turn to the language model history and forcing the decoding to finish in a proper final word.
- Coupled: since the NGT model is used as language model, the previous system turn with the corresponding attached DA labels is added to the current NGT history; additionally, since each turn must present at least a DA label, only hypothesis with a word with an attached DA label are taken into account.

These search restrictions were implemented in the iAtros speech recogniser[7]. iAtros [76] is a speech and handwritten text recogniser based on HMM as acoustic/morphologic models and n-grams or Finite State models as language models. iAtros allows, among other features, the definition of categories and the use of input and output language models for speech translation.

The recognition of user turns employed n-grams language models (with $n = 2, 3, 4, 5$) trained with the SLM toolkit [77], which included the categories that were defined for the task (town names, times, etc.). In the cross-validation experiment, best result was obtained by using a 3-gram, with a Word Error Rate (WER) of 29.1% (with proper words, not categories). This recognised corpus was used in the decoupled turn decoding experiments using the same process that was applied to the transcribed corpus (lowercase, categorisation, etc.).

Coupled decoding, apart from the previously described restrictions, used the iAtros capabilities of using input-output language models (typically used for speech machine translation). In this case, the input language model is the NGT model, were each extended symbol was associated to the acoustic units corresponding to the actual word, and the output language model is the DA n-gram.

In the experiments we tested different values for the degree of the NGT n-gram (in this case, a wider range than in the SwitchBoard case, since preliminary experiments showed that optimal error was obtained for NGT models of higher order) and the DA n-gram (from 2 to 5). The Output Grammar Scale Factor value was optimised in the cross-validation process.

B. Dialogue annotation experiments

Table VI shows the results for the Dihana corpus. The best option is using 4-grams for the NGT model and 3-grams for the DA language model, although differences are not significant with respect to some other combinations of NGT and DA n-gram. In this case, using bigrams in the NGT model or in the DA language model produces poorer results than using higher order n-grams. This could be caused by the task-oriented nature of the Dihana corpus, which implies more regular sequences of words and DA whose span is higher than two words or DA labels.

We can see the high quality in the annotation of the dialogue turns, since more than an 80% of the labels are correct (SegDAER of 17.9, for 4 NGT and 3 DA n-gram). These figures are significantly better than those produced by the HMM/n-grams, CRF and the sequential annotation models (SegDAER 33.7, 26.5, and 22.2 respectively, see Table III, nearly 16%, 9%, and 4% of absolute difference). In this case, since annotation is applied to whole dialogues, both user and system turns are included in the evaluation. NGT provides a segmentation error better than HMM and CRF, and similar to the sequential approach. However, the lower vocabulary size, the regular nature of system turns, and the lower ambiguity of DA labels (due to the task-oriented nature of the corpus, which makes dialogue context more important), make the baseline models (even the SVM of the sequential approach, which does not take into account the context) less accurate than NGT.

DAER and SegER show that, as it happened with SwitchBoard, most of the errors are produced by an incorrect assignment of the DA label, while its position is usually correct. SegER, NIST-SU and DSER show a similar behaviour to that obtained with SwitchBoard. However, SegDAER seems a bit optimistic with respect to *Lenient* and *Strict* measures, but relative behaviour is similar for all the measures.

C. Decoupled turn decoding experiments

In these experiments we used three different versions of the corpora: with and without punctuation marks, and the speech recognised version. The first and second version have the same sense that in SwitchBoard (evaluate NGT with perfect recognition). The speech recognised version corresponds to the output of a real speech recogniser (iAtros), and uses an NGT model trained from dialogues without punctuation marks.

Table VII shows the results for the three versions (with and without punctuation marks, and the speech recognised version) of the Dihana corpus. In this case, results are only for user turns, since system turns are not decoded in a dialogue system (they are produced by the computer, which consequently knows the meaning). The decoding intercalates the corresponding transcription of system turns (with punctuation marks) within the different user turns. To compare the difference with the dialogue-by-dialogue approximation used in annotation, results on annotation for only user turns are shown in the first subtable of Table VII.

In this case, the turn-by-turn approximation produces a significant degradation in DAER with respect to the dialogue-by-dialogue one (from 34.3 to 39.4). Consequently, we can expect a lower performance of the model in the turn decoding task. But in contrast to what happened with the SwitchBoard corpus, the degradation in this case is associated to the lack of information of the development of the dialogue in future turns. This is clear from the comparison between the results with or without punctuation marks, where differences are not significant (from 39.4 to 39.1). The high dependence on future turns could be explained by the more regular nature of DA sequences in this corpus, since it is a task oriented corpus where the interactions follow usually the same order (e.g., asking for times, clarifying data, asking for fares).

In the case of the speech recognised corpus the degradation of the results is quite significant (around 12% absolute DAER), which was expected given that the WER of the speech recog-

[7]iAtros can be downloaded from `www.prhlt.upv.es/page/projects/multimodal/idoc/iatros`.

TABLE VI
DIHANA ANNOTATION RESULTS. FIRST COLUMN REFERS TO THE DEGREE OF THE DA n-GRAM AND SECOND ROW TO THE DEGREE OF THE NGT MODEL. CONFIDENCE INTERVALS ARE LOWER THAN ± 0.7 FOR ALL MEASURES AND EXPERIMENTS.

		NIST-SU				DSER				SegER			
	DA n-gram	2	3	4	5	2	3	4	5	2	3	4	5
NGT	2	5.8	5.7	5.8	6.0	10.1	9.8	10.1	10.5	5.8	5.6	5.8	6.0
	3	4.4	4.3	4.5	4.6	7.3	7.1	7.5	7.6	4.1	4.0	4.1	4.2
	4	4.3	**4.2**	**4.2**	4.3	7.1	**6.9**	**6.9**	7.0	4.0	**3.9**	**3.9**	4.0
	5	4.3	**4.2**	4.3	4.3	7.2	**6.9**	7.0	7.0	4.0	**3.9**	**3.9**	4.0
	6	4.4	4.3	4.4	4.5	7.2	7.0	7.1	7.3	4.2	4.1	4.2	4.2
	7	4.6	4.4	4.6	4.6	7.5	7.2	7.4	7.6	4.3	4.2	4.3	4.4

		Lenient				*Strict*				SegDAER				DAER			
	DA n-gram	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
NGT	2	26.0	24.1	24.6	25.9	26.9	25.1	25.6	26.9	27.5	26.1	26.6	27.7	27.0	25.6	26.1	27.2
	3	20.5	19.1	19.7	20.6	21.5	20.1	20.8	21.6	20.0	18.7	19.4	20.0	19.5	18.2	18.9	19.5
	4	19.5	**18.2**	**18.2**	19.0	20.6	**19.2**	19.3	20.0	19.0	**17.9**	18.1	18.7	18.5	**17.3**	17.5	18.1
	5	20.0	**18.2**	18.3	19.1	21.0	**19.2**	**19.2**	20.1	19.6	18.2	18.4	19.0	19.0	17.6	17.8	18.4
	6	19.4	18.4	18.3	19.0	20.4	19.4	19.3	20.0	19.3	18.5	18.8	19.3	18.7	17.9	18.1	18.6
	7	19.9	19.3	19.1	19.9	21.0	20.3	20.2	20.9	20.5	19.8	19.9	20.5	19.8	19.0	19.1	19.8

TABLE VII
DIHANA DECOUPLED DECODING RESULTS (DAER). CONFIDENCE INTERVALS ARE LOWER THAT ± 1.2 IN ALL EXPERIMENTS.

		Annotation user turns				With punctuation marks				Without punctuation marks				Speech recogniser output			
	DA n-gram	2	3	4	5	2	3	4	5	2	3	4	5	2	3	4	5
NGT	2	46.4	45.2	45.8	47.0	54.8	54.7	54.7	55.1	53.0	52.4	52.9	53.7	65.5	65.0	65.2	65.6
	3	36.7	35.3	36.4	37.4	40.7	39.9	40.3	40.6	40.4	40.0	40.1	40.3	58.1	57.5	57.6	58.4
	4	35.6	**34.3**	34.6	35.5	40.7	40.1	40.0	40.4	39.8	39.2	**39.1**	39.9	58.3	57.5	57.6	58.1
	5	36.7	35.1	35.4	36.3	40.6	39.5	39.6	40.0	40.1	39.2	39.2	40.0	57.8	57.4	57.4	58.0
	6	37.0	36.0	36.5	37.3	40.3	**39.4**	**39.4**	39.7	40.0	39.3	**39.1**	40.1	52.4	**51.5**	51.7	52.3
	7	39.8	38.3	38.5	39.6	41.5	40.6	40.6	41.1	41.3	40.9	40.8	41.7	53.5	52.7	52.9	53.4

TABLE VIII
DIHANA COUPLED DECODING RESULTS (DAER), CONFIDENCE INTERVALS ARE LOWER THAN ± 1.1 IN ALL CASES.

	DA n-gram	2	3	4	5
NGT	2	62.5	62.4	62.4	62.4
	3	57.9	57.8	57.8	57.7
	4	55.2	55.2	55.2	55.2
	5	54.5	54.4	54.4	54.3
	6	54.2	54.1	54.2	54.2
	7	54.2	54.2	54.2	54.2
	8	53.6	**53.5**	**53.5**	**53.5**
	9	53.8	53.7	53.7	53.7
	10	54.0	53.9	53.9	53.9

nised corpus is about 29%. In this case, high order NGT n-grams (6-grams) produce significantly better results. This can be caused by the higher span of these models, which can cover more history and therefore be more robust against recognition errors in the last words of the turn.

In conclusion, for a more usual dialogue system framework (task-oriented, with a more limited vocabulary), the main limitation of the NGT technique is given by the noisy input sequence (produced by the speech recognition engine) and by the local (turn-by-turn) decoding (since for perfect transcriptions decoding results present a significant degradation with respect to annotation results).

D. Coupled turn decoding experiments

The coupled turn decoding experiments used the same cross-validation approach that the decoupled decoding and employed the same acoustic models. Results on coupled decoding for Dihana can be seen in Table VIII. Since tendency of results till NGT of degree 7 showed a reduction of DAER, complementary experiments with higher order NGT models (8 to 10) were performed and their results were included.

As can be seen, differences between using 4 to 10-grams in the NGT model are not significant, but lower order models produce significantly worse results. The use of different DA n-grams does not produce significant differences. When comparing these results with those of the decoupled approximation (Table VII), best results of coupled approximation are a bit worse in absolute terms (53.5 in coupled and 51.5 in decoupled), although these differences are not statistically significant (95% confidence intervals are 53.5 ± 1.1 and 51.5 ± 1.1). Thus, the coupled approximation works as well as the decoupled approximation, but with the advantage of using less time: the decoupled approximation uses about 4.1 seconds per turn for speech decoding and about 0.8 seconds per turn in DA decoding (in total, about 4.9 seconds per turn), while the coupled approximation employs about 1.9 seconds per turn (all times taken in a Intel Core i7 3.4GHz CPU).

IX. CONCLUSIONS AND FUTURE WORK

In this article we presented an extensive description of the NGT technique and its applications to dialogue annotation and turn decoding. Results on dialogue annotation show that the technique is competitive with respect to other more standard approaches. Besides, in a task-oriented dialogue corpus produces better results than for non-task-oriented corpora.

Results on turn decoding for reference transcriptions show that using that approximation could be a valid alternative for

dialogue annotation. The effect of erasing punctuation marks does not produces significant degradation in the task-oriented corpus; therefore, the absence of these marks in the speech decoding must not affect the quality of the DA decoding.

Finally, the decoupled and coupled DA decoding for real speech recognition show no significant differences at DAER level, whereas the decoding time is dramatically lower in the coupled approach with respect to the decoupled one. Therefore, we can conclude that the NGT technique is suitable for performing coupled DA decoding in real dialogue systems, by applying the NGT search in the speech decoding process.

Future work could be directed to improve the features of the speech recogniser, since recognition WER is quite high (29%) and affects severely the final DAER. Results on real transcriptions show a relative DAER increment of a 30% that is caused by this high WER, and show that there is still room for improvement by obtaining better models for speech recognition. Other parameters for the NGT search could be optimised, such as beam search. Another possibility is a combination of models, i.e., using not only NGT but other models such as HMM, Bernoulli distributions or multimodal distributions in order to improve the DA annotation and decoding accuracy. Finally, the implementation of confidence measures on the DA output (for example, by using word-graph-based confidence measures usually employed in speech recognition [78]) would be necessary to eventually give to a dialogue manager more information on the uncertainty of the user input, thus it can react properly by using its dialogue strategy.

ACKNOWLEDGMENTS

Work partially supported by the Spanish MEC/MICINN under STraDA (TIN2012-37475-C02-01) and Active2Trans (TIN2012-31723) projects, and by the Generalitat Valenciana under project AMIIS (ISIC/2012/004). Authors wish to thank the anonymous reviewers for their comments and suggestions.

REFERENCES

[1] Y. Wilks, "Companions: Intelligent, persistent, personalised interfaces to the internet." 2006, http://www.companions-project.org.

[2] P. Trahanias, "Indigo: Interaction with personality and dialogue enabled robots," 2007, http://www.ics.forth.gr/indigo/.

[3] O. Lemon, "The classic project: Computational learning in adaptive systems for spoken conversation," 2011, http://www.classic-project.org/.

[4] H. Hastie, "Parlance: Probabilistic adaptive real-time learning and natural conversational engine," 2012, https://sites. google. com /site /parlanceprojectofficial/home.

[5] W. Wahlster, *SmartKom: Foundations of Multimodal Dialogue Systems (Cognitive Technologies)*. Springer-Verlag New York, Inc., 2006.

[6] V. Petukhova and H. Bunt, "The coding and annotation of multimodal dialogue acts," in *Proc. of LREC'12*. Istanbul, Turkey: ELRA, 2012.

[7] B. J. Grosz, "The structure of task-oriented dialogs," in *Proc. of the IEEE Symposium on Speech Recognition*, Pittsburgh, PA, 1974, pp. 250–253.

[8] K. E. Lochbaum, "A collaborative planning model of intentional structure," *Comput. Linguist.*, vol. 24, no. 4, pp. 525–572, Dec. 1998.

[9] S. Bangalore, G. Di Fabbrizio, and A. Stent, "Learning the structure of task-driven human-human dialogs," *Trans. Audio, Speech and Lang. Proc.*, vol. 16, no. 7, pp. 1249–1259, Sep. 2008.

[10] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer, "Dialogue act modelling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 1–34, 2000.

[11] F. Casacuberta, E. Vidal, and D. Picó, "Inference of finite-state transducers from regular languages," *Pattern Recognition*, vol. 38, no. 9, pp. 1431–1443, 2005.

[12] V. Tamarit, C.-D. Martínez-Hinarejos, and J.-M. Benedí, *Spoken Dialogue Systems Technology and Design*. Springer, 2011, ch. On the Use of N-gram Transducers for Dialogue Annotation, pp. 255–276.

[13] K. Takano and A. Shimazu, "Recognizing local dialogue structures and dialogue acts," in *LPSS'06*, 2006, pp. 263–274.

[14] B. Ludwig, G. Grz, and H. Niemann, "Modelling users, intentions, and structure in spoken dialog," *CoRR*, vol. cs.CL/9809022, 1998.

[15] S. Varges, G. Riccardi, and S. Quarteroni, "Persistent information state in a data-centric architecture," in *Proc. of 9th SIGdial*, 2008, pp. 68–71.

[16] C. Lee, S. Jung, K. Kim, and G. G. Lee, "Hybrid approach to robust dialog management using agenda and dialog examples," *Comput. Speech Lang.*, vol. 24, no. 4, pp. 609–631, Oct. 2010.

[17] D. Jurafsky, R. Ranganath, and D. McFarland, "Extracting social meaning: identifying interactional style in spoken conversation," in *Proc. of HLT-NAACL 2009*. ACL, 2009, pp. 638–646.

[18] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proc. of ICASSP*, 1996, pp. 405–408.

[19] V. Rieser and O. Lemon, "Learning effective multimodal dialogue strategies from wizard-of-oz data: bootstrapping and evaluation," in *Proceedings of ACL-08: HLT*, 2008, pp. 638–646.

[20] J. Schatzmann and S. Young, "The hidden agenda user simulation model," *Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 4, pp. 733–747, May 2009.

[21] S. Young, M. Gašić, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson, and K. Yu, "The hidden information state model: A practical framework for pomdp-based spoken dialogue management," *Comput. Speech Lang.*, vol. 24, no. 2, pp. 150–174, Apr. 2010.

[22] B. Thomson and S. Young, "Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems," *Comput. Speech Lang.*, vol. 24, no. 4, pp. 562–588, Oct. 2010.

[23] F. Jurčíček, B. Thomson, and S. Young, "Reinforcement learning for parameter estimation in statistical spoken dialogue systems," *Comput. Speech Lang.*, vol. 26, no. 3, pp. 168–192, Jun. 2012.

[24] H. Bunt, "Context and dialogue control," *THINK Quarterly*, vol. 3, 1994.

[25] A. Isard, D. McKelvie, and H. Thompson, "Towards a Minimal Standard for Dialogue Transcripts: A New Sgml Architecture for the HCRC Map Task Corpus," in *Proceedings of ICSLP'98*, Sydney, 1998.

[26] S. E. Strayer, P. A. Heeman, and F. Yang, *Reconciling Control and Discourse Structure*. Kluwer, 2003, ch. 14.

[27] M. G. Core and J. F. Allen, "Coding dialogues with the DAMSL annotation scheme," in *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*. AAAI, 1997, pp. 28–35.

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

[29] M. Walker and R. Passonneau, "Date: a dialogue act tagging scheme for evaluation of spoken dialogue systems," in *1st HLT*, 2001, pp. 1–8.

[30] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. ICASSP-92*, 1992, pp. 517–520.

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

[32] H. Hardy, T. Strzalkowski, M. Wu, C. Ursu, N. Webb, A. Biermann, R. B. Inouye, and A. McKenzie, "Data-driven strategies for an automated dialogue system," in *Proc. of 42nd ACL*. ACL, 2004.

[33] R. Prasad and M. Walker, "Training a dialogue act tagger for human-human and human-computer travel dialogues," in *Proc. of 3rd SIGdial*. Stroudsburg, PA, USA: ACL, 2002, pp. 162–173.

[34] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. Traum, "Iso 24617-2: A semantically-based standard for dialogue annotation," in *Proceedings of LREC-2012*. Istanbul, Turkey: ELRA, May 2012, pp. 430–437.

[35] N. Webb, M. Hepple, and Y. Wilks, "Dialogue act classification using intra-utterance features," in *Proceedings of the AAAI Workshop on Spoken Language Understanding*, Pittsburgh, 2005.

[36] C.-D. Martínez-Hinarejos, J.-M. Benedí, and R. Granell, "Statistical framework for a spanish spoken dialogue corpus," *Speech Communication*, vol. 50, pp. 992–1008, 2008.

[37] V. K. R. Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," *Computer Speech and Language*, vol. 23, no. 4, p. 407422, Oct. 2009.

[38] N. Webb and M. Ferguson, "Automatic extraction of cue phrases for cross-corpus dialogue act classification," in *Proceedings of 23rd COLING: Posters*. Stroudsburg, PA, USA: ACL, 2010, pp. 1310–1317.

[39] S. N. Kim, L. Cavedon, and T. Baldwin, "Classifying dialogue acts in one-on-one live chats," in *EMNLP2010*, 2010, pp. 862–871.

[40] K. E. Boyer, J. F. Grafsgaard, E. Y. Ha, R. Phillips, and J. C. Lester, "An affect-enriched dialogue act classification model for task-oriented dialogue," in *Proc. of 49th ACL*. ACL, 2011, pp. 1190–1199.

[41] W.-B. Liang, C.-H. Wu, and C.-P. Chen, "Semantic information and derivation rules for robust dialogue act detection in a spoken dialogue system," in *Proc. of 49th ACL: short papers*. ACL, 2011, pp. 603–608.

[42] T. Klüwer, H. Uszkoreit, and F. Xu, "Using syntactic and semantic based relations for dialogue act recognition," in *Proceedings of 23rd COLING: Posters*. Stroudsburg, PA, USA: ACL, 2010, pp. 570–578.

[43] S. R. Coria and L. A. Pineda, "An analysis of prosodic information for the recognition of dialogue acts in a multimodal corpus in mexican spanish," *Comput. Speech Lang.*, vol. 23, no. 3, pp. 277–310, Jul. 2009.

[44] N. Novielli and C. Strapparava, "Towards unsupervised recognition of dialogue acts," in *HLT-NAACL: Student Workshop*, 2009, pp. 84–89.

[45] G.-A. Levow, "Prosodic cues to discourse segment boundaries in human-computer dialogue," in *Proc. of 5th SIGdial*. ACL, 2004, pp. 93–96.

[46] N. Netten and M. van Someren, "Identifying Segments for Routing Emergency Response Dialogues," in *5th ISCRAM Conference*, 2008.

[47] M. Zimmermann, D. Hakkani-Tr, J. G. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, "The icsi+ multilingual sentence segmentation system." in *INTERSPEECH*. ISCA, 2006.

[48] U. Guz, S. Cuendet, D. Hakkani-Tür, and G. Tur, "Multi-view semi-supervised learning for dialog act segmentation of speech," *Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 2, pp. 320–329, Feb. 2010.

[49] M. Atterer, T. Baumann, and D. Schlangen, "Towards incremental endof-utterance detection in dialogue systems," in *Coling*, 2008, pp. 11–14.

[50] U. Guz, G. Tur, D. Hakkani-Tür, and S. Cuendet, "Cascaded model adaptation for dialog act segmentation and tagging," *Comput. Speech Lang.*, vol. 24, no. 2, pp. 289–306, Apr. 2010.

[51] F. Morbini and K. Sagae, "Joint identification and segmentation of Domain-Specific dialogue acts for conversational dialogue systems," in *Proceedings of 49th ACL*, vol. 2. Portland, OR: ACL, Jun. 2011.

[52] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "A* based joint segmentation and classification of dialog acts in multi-party meetings," in *IEEE ASRU*, 2005, pp. 581–584.

[53] M. Harper, B. Dorr, J. Hale, B. Roark, I. Shafran, M. Lease, Y. Liu, M. Snover, L. Yung, A. Krasnyanskaya, and R. Stewart, "2005 jhsw final report on parsing and spoken structural event detection," 2005.

[54] A. Dielmann and S. Renals, "DBN based joint dialogue act recognition of multiparty meetings," in *Proc of ICASSP '07*, 2007.

[55] A. Stolcke, E. Shriberg, R. A. Bates, M. Ostendorf, D. Hakkani, M. Plauch, G. Tr, and Y. Lu, "Automatic detection of sentence boundaries and disfluencies based on recognized words." in *ICSLP*. ISCA, 1998, pp. 2247–2250.

[56] A. Stolcke, E. Shriberg, D. Z. Hakkani-Tr, and G. Tr, "Modeling the prosody of hidden events for improved word recognition." in *EUROSPEECH*. ISCA, 1999, pp. 307–310.

[57] C.-D. Martínez-Hinarejos, "A study of a segmentation technique for dialogue act assignation," in *8th IWCS*, 2009, pp. 299–304.

[58] J. M. Benedí, E. Lleida, A. Varona, M. J. Castro, I. Galiano, R. Justo, I. López, and A. Miguel, "Design and acquisition of a telephone spontaneous speech dialogue corpus in spanish: Dihana," in *Fifth LREC*, Genova, Italy, 2006, pp. 1636–1639.

[59] N. Webb and Y. Wilks, "Error analysis of dialogue act classification," in *Proceedings of the 8th TSD*, 2005, pp. 451–458.

[60] M. Fraser and G. Gilbert, "Simulating speech systems," *Comp. Speech Lang.*, vol. 5, pp. 81–99, 1991.

[61] T. Fukada, D. Koll, A. Waibel, and K. Tanigaki, "Probabilistic dialogue act extraction for concept based multilingual translation systems," in *Proceedings of ICSLP*, vol. 6, 1998, pp. 2771–2774.

[62] N. Alcácer, J. M. Benedí, F. Blat, R. Granell, C. D. Martínez, and F. Torres, "Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus," in *SPECOM*, Greece, 2005, pp. 583–586.

[63] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *ICASSP '05*, vol. 1, 2005, pp. 1061–1064.

[64] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward joint segmentation and classification of dialog acts in multiparty meetings," in *Proc. of 2nd MLMI*. Springer-Verlag, 2006, pp. 187–193.

[65] C. Sutton and A. McCallum, "An introduction to conditional random fields," *Foundations and Trends in Machine Learning*, vol. 4, no. 4, pp. 267–373, 2012.

[66] S. N. Kim, L. Cavedon, and T. Baldwin, "Classifying dialogue acts in multi-party live chats," in *Proc. of the 26th PACLIC*, 2012, pp. 463–472.

[67] D. Verbree, R. Rienks, and D. Heylen, "Dialogue-act tagging using smart feature selection: results on multiple corpora," in *SLT*, 2006.

[68] G. Tur, U. Guz, and D. Hakkani-Tur, "Model adaptation for dialog act tagging," in *SLT 2006*, Mar. 2006, pp. 94–97.

[69] N. Webb and T. Liu, "Investigating the portability of corpus-derived cue phrases for dialogue act classification," in *COLING*, 2008, pp. 977–984.

[70] V. K. R. Sridhar, S. Bangalore, and S. Narayanan, "Combining lexical, syntactic and prosodic cues for improved online dialog act tagging," *Computer Speech and Language*, vol. 23, no. 4, pp. 407 – 422, 2009.

[71] S. Quarteroni, A. V. Ivanov, and G. Riccardi, "Simultaneous dialog act segmentation and classification from human-human spoken conversations," in *Proc. ICASSP*, May 2011.

[72] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *ICASSP'04*, vol. 1, 2004, pp. 409–412.

[73] G. Tur, "Co-adaptation: Adaptive co-training for semi-supervised learning," in *Proc. of ICASSP*, 2009, pp. 3721–3724.

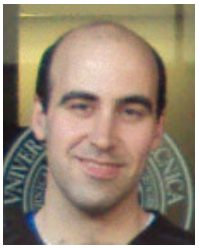
[74] ——, "Extending boosting for large scale spoken language understanding," *Mach. Learn.*, vol. 69, no. 1, pp. 55–74, Oct. 2007.

[75] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. CUED, 2006.

[76] M. Luján-Mares, V. Tamarit, V. Alabau, C. D. Martínez-Hinarejos, M. P. i Gadea, A. Sanchis, and A. H. Toselli, "iatros: A speech and handwritting recognition system." in *VJTH'2008*, 2008, pp. 75–78.

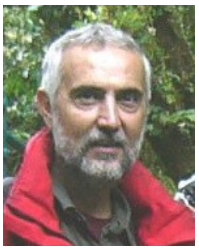
[77] R. Rosenfeld, "The cmu-cambridge statistical language modelling toolkit v2," Carnegie Mellon University, Tech. Rep., 1998.

[78] A. Sanchis, A. Juan, and E. Vidal, "A word-based naïve bayes classifier for confidence estimation in speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 565–574, 2012.



Carlos-D. Martínez-Hinarejos Carlos-David Martínez-Hinarejos obtained his BSc in Computer Science in 1998, his Ph.D. degree in Pattern Recognition and Artificial Intelligence in 2003, and his BSc in Biotechnology in 2012, all from Universitat Politècnica de València (UPV). He joined to the UPV staff (Computation and Computer Systems Department, DSIC) in 2000 as assistant (*Ayudante*), and from 2007 as senior lecturer (*Prof. Contratado Doctor*). In 2005 he visited University of Sheffield, Department of Computer Science, for three months.

Dr. Martínez-Hinarejos pertains to the Pattern Recognition and Human Language Technology (PRHLT) research center, where he develops his research on the topics of speech recognition, dialogue systems, multimodal systems, text classification, and bioinformatics. He has participated in many European and Spanish projects, and is an active member of the Spanish Network for Speech Technologies (RTTH).



José-M. Benedí José Miguel Benedí earned his BSc in Physics from the University of Valencia in 1983. He received his Ph.D in Computer Science from the Polytechnic University of Valencia in 1989.

From 1984 to 1986, he worked in the Department of Electricity and Electronics, University of Valencia, as Assistant Professor. Since 1987, he has been at the Department of Information Systems and Computation, Polytechnic University of Valencia, first as Associate Professor and later, from 2002 onwards, as Full Professor. He has been an active member of the Pattern Recognition and Human Language Technology Group since 1984. His current research interests include the areas of statistical and syntactic pattern recognition, machine learning, and their applications to language, speech, and image processing.

Dr. Benedí is a member of the Spanish Society for Pattern Recognition and Image Analysis (AERFAI), an affiliate society of IAPR, the IEEE Computer Society, and the Spanish Association for Artificial Intelligence (AEPIA).

Vicent Tamarit Vicent Tamarit obtained his BSc in Computer Science in 2006, his MSc in Artificial Intelligence, Pattern Recognition, and Digital Image in 2008, and his BSc in Journalism in 2012. He founded Nosplay.com in 2010 and is currently working as a Public Relations and Marketing manager in Codigames.