Improving Unsegmented Dialogue Turns Annotation with N-gram Transducers*

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

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

Abstract. The statistical models used for dialogue systems need annotated data (dialogues) to infer their statistical parameters. Dialogues are usually annotated in terms of Dialogue Acts (DA). The annotation problem can be attacked with statistical models, that avoid annotating the dialogues from scratch. Most previous works on automatic statistical annotation assume that the dialogue turns are segmented into the corresponding meaningful units. However, this segmentation is not usually available. Most recent works tried the annotation with unsegmented turns using an extension of the models used in the segmented case, but they showed a dramatical decrease in their performance. In this work we propose an enhanced annotation technique based on N-gram transducers that outperforms the accuracy of the classical HMM-based model for annotation and segmentation of unsegmented turns.

Keywords: Dialogue annotation, statistical models

1 Introduction

A dialogue system is usually defined as a computer system that interacts with a human user to fulfil a task (Dybkjær and Minker, 2008). These systems are of particular interest in many applications, like information systems that are accessed by telephone (Seneff and Polifroni, 2000; Aust et al., 1995) or assistant systems for people with special necessities (Wilks, 2006). All the systems define the way they react to user inputs with the so-called dialogue strategy. This strategy can be rule-based (Gorin et al., 1997) or data-based (Young, 2000).

In any case, the strategies are based on the interpretation of the user input in terms of dialogue semantic units. These semantic units are usually coded in terms of Dialogue Acts (DA) (Bunt, 1994), which model the intention of the current user interaction along with its associated information. This concept can be extended to system responses. In an interaction, several dialogue meaningful units can be distinguished. These units are called segments (or utterances according to authors such as (Stolcke et al., 2000)), and each segment has associated only one DA label.

The annotation of a dialogue corpus in terms of DA is an interesting problem for both the development of data-based dialogue systems and the study of discourse and dialogue structure. In the first case, the statistical models that implement the dialogue manager (Williams and Young, 2007; Meng et al., 2003; Stolcke et al., 2000) rely on annotated dialogues to estimate their parameters. This annotation process is developed by human experts and it is a hard and time-consuming task. The use of probabilistic models can provide a draft annotation of the corpus (Stolcke et al., 2000) that can make the manual annotation process faster.

Most of the previous works on the use of probabilistic models for DA annotation use segmented dialogue turns (Stolcke et al., 2000; Webb and Wilks, 2005; Rangarajan et al., 2007). However, this segmentation is not usually available in the initial transcription of a dialogue corpus. Other works propose a decouple segmentation-annotation scheme (Ang et al., 2005), but the ideal option

* Work partially supported by the EC (FEDER) and the Spanish MEC under the MIPRCV "Consolider Ingenio 2010" research programme, the grant TIN2006-15694-CO2-01 and the PROMETEO/2009/014 project.

is the use of models that can annotate unsegmented dialogue turns, giving the correct segments and labels. This option has been explored in a few previous works (Zimmermann et al., 2005; Martínez-Hinarejos et al., 2008), giving in any case (as could be expected) poorer results than when the segmentation is available.

The classical model for this task is based on Hidden Markov Models (HMM) (Stolcke et al., 2000). In this work we present an enhanced version of an alternative model, the N-Gram Transducer (NGT) model. This enhancement provides competitive results with respect to the classical HMM approach in the annotation and segmentation accuracy for unsegmented dialogues, even in two dialogue corpora of very different nature.

This paper is organised as follows: in Section 2 we present the two statistical models that are compared, in Section 3 we detail the corpora for the experiments, in Section 4 we describe the experiments and show their results, in Section 5 we draw conclusions and future lines of work.

2 Statistical annotation models

In this section we present the statistical annotation models that we are going to compare: a classical HMM-based model and the enhanced NGT model. Both models are oriented to solve the optimisation problem of, given a word sequence $\mathcal{W}$ that represents a dialogue, obtaining the sequence of DA labels $\mathcal{U}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$. We can express the complete sequences of words and DA in terms of the different turns in the dialogue: given a dialogue with T turns, we express its associated word sequence and DA sequence as $\mathcal{W} = W_1^T = W_1 W_2 \cdots W_T$ and $\mathcal{U} = U_1^T = U_1 U_2 \cdots U_T$, respectively. Thus, we can express the optimisation problem as:

$$\widehat{\mathcal{U}} = \operatorname*{argmax}_{\mathcal{U}} \Pr(\mathcal{U}|\mathcal{W}) = \operatorname*{argmax}_{U_1^T} \Pr(U_1^T|W_1^T) \tag{1}$$

From Equation 1, we can develop a model based on the application of the rule of Bayes on the formula (HMM-based model) or a model based on the direct implementation of the posterior probability (NGT model).

2.1 The HMM-based model

Using the rule of Bayes, Equation 1 can be expressed as:

$$\operatorname*{argmax}_{U_1^T} \Pr(U_1^T|W_1^T) = \operatorname*{argmax}_{U_1^T} \Pr(U_1^T) \Pr(W_1^T|U_1^T) = \operatorname*{argmax}_{U_1^T} \prod_{t=1}^{T} \Pr(U_t|U_1^{t-1}) \Pr(W_t|W_1^{t-1}, U_1^T) \tag{2}$$

Previous works such as that presented in (Stolcke et al., 2000) have proposed similar approaches to DA annotation. However, these previous approximations assume the availability of the segmentation of the turn to perform the DA assignation (Stolcke et al., 2000; Webb and Wilks, 2005), when segmentation of the turns is not usual in transcribed dialogues. In our case, we try to generalise the DA assignation problem in the case of unavailable segmentation.

We can develop the formulation to use the model in the unsegmented case as[1]: given the current word sequence $W_t = w_1^l = w_1 w_2 \cdots w_l$, we describe it in terms of all the possible segmentations as $W_t = w_{s_0+1}^{s_1} w_{s_1+1}^{s_2} \ldots w_{s_{r-1}+1}^{s_r}$, where r is the number of segments and s_k is the index of the k-th segment. Moreover, we can express the DA sequence of turn t as $U_t = u_1^r$ and the previous DA sequences as $U_1^{t-1} = U_1 U_2 \cdots U_{t-1}$. Furthermore, as W_1^T is the sequence of given events, we can neglect the dependency between word sequences. We can reasonably assume that

[1] Notice that w and u represent single words and DA, respectively, while W and U represent turns and DA sequences.

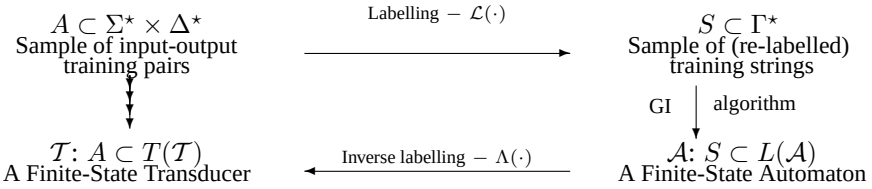

Figure 1: General scheme for the GIATI technique. Σ, Δ and Γ are the input, output, and extended set of symbols, respectively. A and S are the initial sets of aligned and re-labelled samples. $L(\mathcal{A})$ and $T(\mathcal{T})$ represent the languages derived from $\mathcal{A}$ and $\mathcal{T}$, respectively.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

DA sequences until turn t only affect the first t turns in the dialogue (i.e., $\Pr(W_t|W_1^{t-1}, U_1^T) \approx \Pr(W_t|W_1^{t-1}, U_1^t)$). Consequently, the terms in the product in Equation 2 are rewritten as:

$$\Pr(U_t|U_1^{t-1}) \Pr(W_t|W_1^{t-1}, U_1^t) \approx \sum_{r, s_1^r} \prod_{k=1}^{r} \Pr(u_k|u_1^{k-1}, U_1^{t-1}) \Pr(w_{s_{k-1}+1}^{s_k}|u_1^k, U_1^{t-1}) \qquad (3)$$

This model in Equation 3 can be simplified with some assumptions: the current DA depends only on the previous $n-1$ DA and the sequence of words of the current segment depends only on the current DA. The search problem given by this model (the search for the segmentation and DA sequence with maximum probability) is solved using the Viterbi process, which implies that the summation is changed by a maximisation. Therefore, the final model is:

$$\widehat{\mathcal{U}} = \underset{\mathcal{U}}{\operatorname{argmax}} \prod_{t=1}^{T} \max_{r, s_1^r} \prod_{k=1}^{r} \Pr(u_k|u_{k-n+1}^{k-1}) \Pr(w_{s_{k-1}+1}^{s_k}|u_k) \qquad (4)$$

Following the work of other authors (Stolcke et al., 2000; Young, 2000), the terms in Equation 4 are modelled as follows: $\Pr(u_k|u_{k-n+1}^{k-1})$ is usually represented by a statistical model of DA sequences (DA language model), generally an n-gram model, and $\Pr(w_{s_{k-1}+1}^{s_k}|u_k)$ is usually modelled by a HMM. This formulation searches, using a Viterbi process, for the DA sequence of a complete dialogue and gives as by-product a segmentation for each turn. In case there is an available segmentation, the maximisation step is overridden and the values r and s_1^r are fixed to that provided by the segmentation. In any case, the influence of the models (specially the DA language model) can be tuned by using scaling factors, similar to the Grammar Scale Factor used in speech recognition.

2.2 The NGT model

The alternative model is the NGT model, which directly estimates the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$ by means of an n-gram model which acts as a transducer. The definition of this model is based on a Stochastic Finite-State Transducer (SFST) inference technique known as GIATI[2] (Casacuberta et al., 2005). GIATI is a general technique to infer SFST whose first application was in Machine Translation. GIATI starts from a corpus of aligned pairs of input-output sequences. These alignments are used in a re-labelling process that produces a corpus of extended words as a result of a combination of the words of the input and output sentences. This corpus is used to infer a grammatical model (usually a smoothed n-gram). The inversion of the re-labelling process on the grammatical model results in the final SFST, although the use of smoothing techniques makes difficult the conversion of the n-gram to an equivalent SFST. The general GIATI process is presented in Figure 1.

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

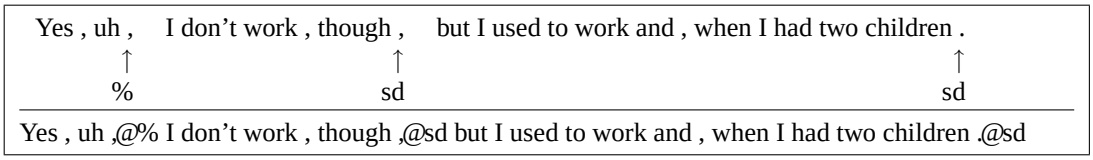

Figure 2: An alignment between a dialogue turn and its corresponding DA labels (from the SWBD-DAMSL scheme, %: uninterpretable, sd: statement-non-opinion), and the result of the re-labelling process, where @ is the attaching metasymbol.

In the case of dialogues, the input language is the sequence of words of the dialogue $w_1 w_2 \dots w_l$, the output language is the sequence of DA of the dialogue $u_1 u_2 \dots u_r$, and the alignment is between the last word of the segment and the corresponding DA. The re-labelling step attaches the DA label to the last word of the segment using a metasymbol (@), providing the extended word sequence $e_1 e_2 \dots e_l$, where $e_i = w_i$ when w_i is not aligned to any DA and $e_i = w_i @ u_k$ 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 the re-labelling process, a grammatical model is inferred. The usual option is a smoothed n-gram.

In the case of dialogues, we can avoid the conversion to SFST. The alignments between the words in the turn and the corresponding DA labels are monotonic (no cross-inverted alignments are possible), and consequently no conversion to SFST is necessary to efficiently apply a search algorithm on the n-gram. Therefore, this n-gram acts as a transducer and gives the name to the technique (NGT: N-Gram Transducers) (Martínez-Hinarejos et al., 2008).

The decoding in the NGT model is a Viterbi search in which each input word is expanded for all the possible outputs it has associated in the alignments in the training corpus. Therefore, the tree search is expanded in each node in several branches according to the number of outputs associated to the word. Each new branch represents a possible output (DA), including the empty output (the word is not attached to any DA).

The probability of each branch is updated according to the corresponding n-gram probability. We start from a parent node P associated to the sequence of extended words $e_1 e_2 \dots e_{i-1}$ and with an associated probability p_P. If the new word to process w_i has associated in the training corpus o outputs and the empty output, we expand from P the children nodes $e_i^0 = w_i$, $e_i^1 = w_i @ u^1$, ..., $e_i^o = w_i @ u^o$. The probability of the child node associated to the extended word e_i^j is computed as $p_P \cdot \Pr(e_i^j | e_{i-n} \dots e_{i-1})$, where $\Pr(e_i^j | e_{i-n} \dots e_{i-1})$ is given by the n-gram of the NGT model.

An example of tree expansion is presented in the top tree of Figure 3. This expansion on complete dialogues produces a high temporal and spatial complexity, which is admissible in the off-line dialogue annotation framework. The search process can be applied to dialogues with unsegmented turns giving, as in the case of the HMM-based model, a segmentation as by-product. NGT can be applied on segmented turns by restricting the outputs to the end words of the segments.

The main drawback of this initial approach is its high locality: only the last n extended words are really taken into account to assign the DA labels. This makes the current DA independent from most of the previous DA in the dialogue history, and loses an important source of information. In the top tree of Figure 3 we can see that the last node of the best hypothesis (in boldface and marked by an arrow) calculates its probability based only on the two previous nodes values ("don't work"), ignoring previous DA. We propose a modification of the basic search algorithm in which the probability of the different branches is not only computed from the n-gram transducer itself, but from an n-gram of DA as well (which acts as DA language model). Therefore, when expanding a branch of a word w_i with a DA u_j, the new probability is computed using both the n-gram transducer and the DA language model (a n-gram of degree m). Consequently, the probability for the child node associated to $e_i = w_i @ u_j$ is given by $p_P \cdot \Pr(e_i | e_{i-n} \dots e_{i-1}) \cdot \Pr(u_j | u_{j-m} \dots u_{j-1})$. No change in the computation of the probability of the child node is produced when the output is empty (i.e.,

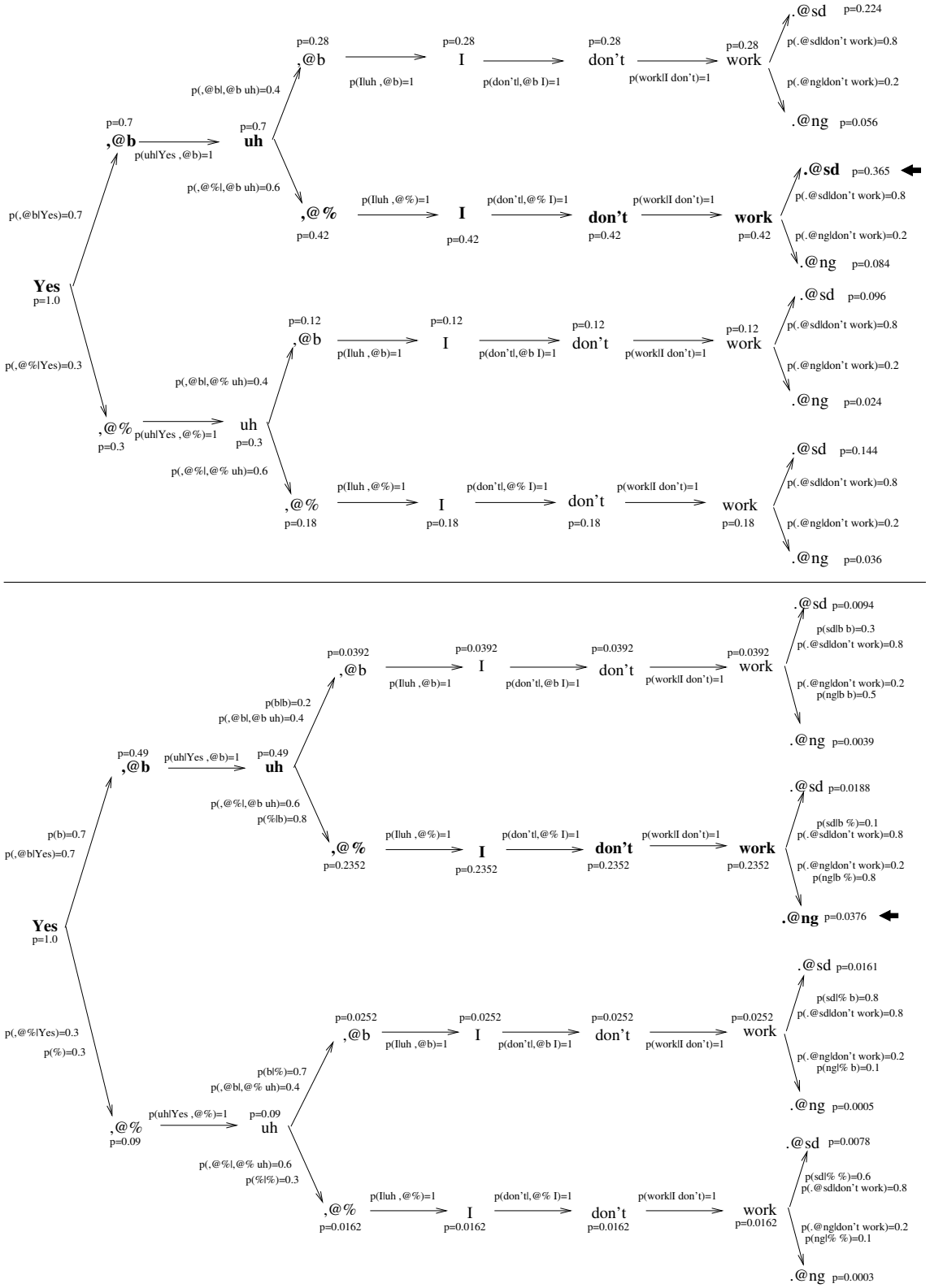

Figure 3: An example of the Viterbi tree search for the basic NGT model (top) and for the enhanced NGT model (bottom). In the basic model only the NGT probabilities are taken into account. In contrast, in the enhanced model both the NGT probabilities and the DA language model probabilities are used in the transitions where outputs are produced. In these nodes, the new probability is the result of multiplying the probabilities of the parent node, the NGT probability and the DA language model probability. Best hypothesis (in boldface and marked by the dark arrow) changes from (b % sd) to (b % ng). In this example, trigrams are used in all models.

$e_i = w_i$). The expansion process in this new version can be seen in the bottom tree of Figure 3. We can see that the probability of the last node of the best hypothesis is calculated using the values of the two previous nodes ("don't work") and the two previous DA ("b %").

With this enhancement, the NGT model keeps information on the DA history and is competitive with respect to the HMM-based model, as the results in Section 4 will show.

3 Corpora

In this section we present the corpora on which we carried out the experiments to compare the HMM-based model and the NGT model. These two corpora are of very different nature, allowing us to generalise the conclusions obtained from the results of the experiments.

3.1 SwitchBoard corpus

The SwitchBoard corpus (Godfrey et al., 1992) is a well-known corpus of human-to-human telephonic conversations in English. The conversations are about general topics, with no clear task to accomplish. This corpus recorded spontaneous speech, with frequent interruptions between the speakers, hesitations, non-linguistic sounds (laugh, cough) and background noises. The transcription of the corpus takes these phenomena into account, and it includes special notation for the overlaps and different noises produced in the recording.

The corpus consists of 1,155 conversations, with approximately 115,000 different turns. The vocabulary size is about 42,000 words. The dialogue annotation was performed using the SWBD-DAMSL scheme (Jurafsky et al., 1997), a simplified version of the standard DAMSL annotation set (Core and Allen, 1997). In the process, the dialogue turns were split into segments and each segment was annotated with one of the 42 different labels of the SWBD-DAMSL scheme. These labels represent several dialogue communicative categories such as statement, question, backchannels, etc., and the corresponding subcategories (e.g., statement opinion/non-opinion, yes-no/open question, etc.). The manual labelling was performed by 8 different human labellers, with a Kappa value of 0.80 (Stolcke et al., 2000).

To simplify the experimental framework, we preprocessed the SwitchBoard corpus to remove certain phenomena: interruptions and overlaps were erased (by joining the interrupted turns), all the words were transcribed to lowercase and punctuation marks were separated from the words. This preprocess is reasonable for the annotation of transcribed dialogues, but for speech dialogues it should be changed (as punctuation marks are not usually part of speech recognisers outputs).

3.2 Dihana corpus

The Dihana corpus (Benedí et al., 2006) is a set of 900 task-oriented human-computer dialogues in Spanish. The task is about railway information for timetables, fares and services for long-distance trains in the Spanish territory. The corpus was acquired from conversations with 225 voluntary speakers, with small Spanish dialectal variants. The acquisition was performed using the Wizard of Oz technique (Fraser and Gilbert, 1991), and it only had semantic restrictions (the objective of the interaction was defined by mean of scenarios), but not lexical or syntactical restrictions.

The acquisition process resulted in 6,280 user turns and 9,133 system turns, with a vocabulary of approximately 900 words and a final amount of speech signal of about five and a half hours. On average, there are 15 words and 1.5 segments per turn. The dialogue annotation scheme was defined based on the Interchange Format (IF) (Fukada et al., 1998), which defines labels with three different levels, called respectively speech act, concept and argument. The adaptation for the Dihana corpus resulted in a set of 248 different 3-level labels (153 for user turns and 95 for system turns) (Alcácer et al., 2005). Due to the high specificity of the third level (which takes into account the specific data used or provided in the segment), an alternative labelling using only the first two levels is also considered in the experiments. In this 2-level case, there are 72 different labels (45 for user and 27 for system).

To simplify the experimental framework, the Dihana corpus was preprocessed to reduce its complexity. In this case, as in the case of the SwitchBoard corpus, all the words were transcribed to lowercase and punctuation marks were separated from the words. Additionally, a categorisation of sequences such as town names, dates, hours, etc. was performed, and the words were speaker-labelled (U for user and S for system).

4 Experiments and results

We propose a set of experiments to compare the performance of the two models introduced in Section 2. The models were proved with the two corpora described in Section 3, and the experiments were made using a cross-validation approach. For both corpora, we present results for the annotation using the segmented and unsegmented version of the dialogues. In both cases the weight parameter of the HMM-based model (which scales the influence of the DA language model) was optimised for the whole cross-validation process. We only show the results for the best weight parameter. The Viterbi search in the NGT model was a beam-search with a dynamic beam parameter. We used the following evaluation metrics:

- Segmented version: CER (Classification Error Rate), i.e., percent of segments with incorrect DA assignation; it is the lower bound of error, as no insertion/deletions are possible.
- Unsegmented versions: different types of Error Rates (ER) based on average edit distance [3] between the reference and the annotation result; the different measures are DAER (Dialogue Act ER) for DA sequences, SegER (Segmentation ER) for segmentation (end positions of the segments) and SegDAER, where the symbols that are compared in the edit distance join both the DA label and its position (segmentation), giving a joining measure of the precision of both the DA assignation and the segmentation.

Although other evaluation metrics can be used (see (Ang et al., 2005)), we consider these metrics a good choice to evaluate the quality of the techniques in the annotation and segmentation task. In all the experiments, confidence intervals of 90% were calculated using bootstrapping with 1000 repetitions using the method described in (Bisani and Ney, 2004).

4.1 SwitchBoard experiments

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

Table 1 shows the results of the annotation with the HMM-based and NGT models with transducer n-grams of degrees 3 and 4 (which were those that offered best results in the overall test experiments). We included the error measures for different estimations of the DA language model ($\Pr(u_k|u_{k-n+1}^{k-1})$) using the segmented and unsegmented version of the corpus. The results in the segmented case are similar to those reported by other authors (Stolcke et al., 2000; Webb and Wilks, 2005; Rangarajan et al., 2007), although not directly comparable due to the different experimental framework.

As was expected, the availability of the segmentation allows a better annotation of the dialogues (for the two techniques, see top left subtable of Table 1). The HMM-based method produces the best results in the annotation of the segmented version. However, in the more realistic and complex unsegmented case, the NGT model is significantly better than the HMM-based model in pure annotation (8% less absolute DAER, including confidence intervals, bottom left subtable of Table 1) and specially when segmentation is included in the evaluation (10% less SegDAER, bottom right subtable of Table 1). This could be caused by the nature of the HMM models, which are one-state models as they have to model the shortest segments (one word) and, consequently, cannot discriminate appropriately words that mark the end of a segment.

[3] $(ins + del + sub)/(ok + del + sub)$, with ins insertions, del deletions, sub/ok wrong/correct substitutions.

Table 1: Annotation and segmentation errors of the SwitchBoard corpus with different n-grams to estimate the DA language model. The first line for each subtable corresponds to the HMM method and the next ones stand for the NGT method with different degrees of the transducer n-gram. CER (top left) corresponds to the segmented case, and the rest of measures to the unsegmented case. 90% confidence intervals are ≤ 0.2 in all cases. In boldface, the best result for each method.

		DA language model			
	Method	2	3	4	5
CER	HMM	**34.4**	34.5	35.0	36.1
	NGT 3g	40.6	38.9	**38.8**	39.2
	NGT 4g	41.8	40.3	39.8	40.3
DAER	HMM	**54.6**	55.5	55.8	57.0
	NGT 3g	46.7	**46.6**	46.7	47.0
	NGT 4g	48.3	48.0	48.1	48.3

		DA language model			
	Method	2	3	4	5
SegER	HMM	**41.2**	41.8	41.7	42.3
	NGT 3g	**22.9**	**22.9**	**22.9**	**22.9**
	NGT 4g	24.0	24.0	23.9	24.0
SegDAER	HMM	**60.4**	61.8	62.0	63.6
	NGT 3g	**50.4**	**50.4**	50.5	50.8
	NGT 4g	52.6	52.4	52.5	52.7

Table 2: Annotation and segmentation errors of the Dihana corpus using the 2-level labelling. Different n-grams to estimate the DA language model are proved. The first line of each subtable corresponds to the HMM method and the next ones stand for the NGT method with different degrees of the transducer n-gram. CER (top left) corresponds to the segmented case, and the rest of measures to the unsegmented case. 90% confidence intervals are in all cases ≤ 0.5. In boldface, the best result for each method.

		DA language model			
	Method	2	3	4	5
CER	HMM	6.2	**5.8**	5.8	6.3
	NGT 3g	9.7	7.9	7.8	8.1
	NGT 4g	8.5	**7.3**	7.3	7.4
DAER	HMM	9.0	**7.9**	8.1	8.5
	NGT 3g	10.6	8.6	8.6	8.7
	NGT 4g	9.2	**7.9**	8.0	8.1

		DA language model			
	Method	2	3	4	5
SegER	HMM	23.1	**22.9**	23.0	23.0
	NGT 3g	1.2	1.3	1.3	1.2
	NGT 4g	**1.1**	**1.1**	**1.1**	**1.1**
SegDAER	HMM	30.4	**29.2**	29.4	29.7
	NGT 3g	11.1	9.1	9.0	9.2
	NGT 4g	9.5	**8.3**	**8.3**	8.5

In the segmented version, the annotation error is slightly affected by the DA language model degree, but these differences are not significant in the unsegmented case. This can be caused by the nature of the SwitchBoard corpus, as the dialogues are not task-oriented. Consequently, the DA sequences do not follow regular patterns that can be captured by the DA language models.

4.2 Dihana experiments

The experiments were also performed with the Dihana corpus. The corpus was divided into 5 partitions to carry out a cross-validation approach. Each partition contains 180 dialogues. We applied the models to annotate the corpus with the 2-level and the 3-level labels.

Table 2 shows a comparison of the two annotation methods using the 2-level labels. The tests were made using the segmented and unsegmented versions of the corpus. With the 2-level annotation, the HMM-based model is the best annotation method for the segmented and the unsegmented versions. However, the segmentation is significantly worse than that produced by NGT (see SegER and SegDAER results, left subtables of Table 2), with an error difference higher than 20%. Thus, NGT seems more convenient in the annotation task, where the correct boundaries are as important as correct DA labels.

The experiments with the 3-level labels are shown in Table 3. The HMM-based model is, with this set of labels, significantly better in pure annotation in both versions of the corpus. This may be due to the number of 3-level labels, that produces a DA language model with higher perplexity. This affects specially to the NGT model because this model only uses the last n words of the segment to obtain the label hypothesis, while the HMM-based model uses the whole word sequence

Table 3: Annotation and segmentation errors of the Dihana corpus using the 3-level labelling. Different n-grams to estimate the DA language model are tested. The first line in each subtable corresponds to the HMM method and the next ones stand for the NGT method with different degrees of the transducer n-gram. CER (top left) corresponds to the segmented case, and the rest of measures to the unsegmented case. 90% confidence intervals are in all cases ≤ 0.6. In boldface, the best result for each method.

		DA language model			
	Method	2	3	4	5
CER	HMM	10.8	**10.3**	10.5	11.0
	NGT 3g	18.2	16.9	17.2	17.9
	NGT 4g	17.1	**16.0**	16.2	16.7
DAER	HMM	**15.3**	15.6	15.9	16.7
	NGT 3g	19.5	18.2	18.9	19.5
	NGT 4g	18.4	**17.4**	17.5	18.1

		DA language model			
	Method	2	3	4	5
SegER	HMM	**25.6**	26.2	26.2	26.4
	NGT 3g	4.2	4.1	4.3	4.4
	NGT 4g	4.1	**4.0**	**4.0**	4.1
SegDAER	HMM	**33.6**	34.5	34.7	35.4
	NGT 3g	20.0	18.7	19.4	20.0
	NGT 4g	19.0	**17.9**	18.1	18.7

of the segment. However, when segmentation is included in the evaluation, the performance of the HMM-based model dramatically decreases (more than 15%) with respect to the NGT model.

The low accuracy on the segmentation of the HMM-based model is caused by the same reasons that were pointed out for the SwitchBoard corpus (one-state HMM models that cannot correctly capture the boundary words). The influence of the DA language model is slightly more important in the Dihana corpus than in the SwitchBoard corpus. This difference is due to the nature of the corpus: Dihana is a task-oriented corpus with a limited number of scenarios; this causes dependencies between the DA sequences that are captured by the DA language model. Consequently, the DA language model degree has a higher influence in the annotation and segmentation.

5 Conclusions and future work

In this work, we presented the enhanced NGT model for dialogue annotation. This model was compared with the classical HMM-based model for dialogue annotation and segmentation using two corpora. The experiments showed that the NGT model produces better error rates in segmentation and annotation with respect to the HMM-based model for two different dialogue corpora.

In pure annotation, the HMM-based model showed a better performance in the segmented dialogues. This could be explained by the use of whole sequence of words and the unambiguity of the segmentation, whereas the NGT model only takes into account the final words of the segment. However, the HMM-based model seems to be more sensitive to the lack of the correct segmentation than the NGT model. Moreover, segmentation results of the HMM-based model are poorer than those of the NGT model. Consequently, the NGT model seems more appropriate for the annotation of unsegmented dialogue corpora.

The NGT model is based on an n-gram inferred from a previously annotated corpus. This n-gram can be used directly in a speech recognition system as language model, producing a DA annotated recognition. Consequently, future work is directed to include this annotation model into a speech recognition system. This needs the development of an NGT model for turn-by-turn annotation, instead of the whole dialogue annotation model presented in this paper.

In the NGT model the search process can be modified with a scale factor that affects the frequency of outputs, or even the DA language model (similar to the scale factor of the HMM-based model). Therefore, another interesting work is the implementation of the search algorithm for the NGT model with these factors and a study of their influence in the annotation results. Finally, we plan to study the efficiency of the HMM-based and NGT models and improve the time complexity of the search process (e.g., using beam-search or N-best expansion).

References

Alcácer, N., J. M. Benedí, F. Blat, R. Granell, C. D. Martínez, and F. Torres. 2005. Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. In *SPECOM*, pages 583–586, Greece.

Ang, J., Y. Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the ICASSP*, volume 1, pages 1061–1064, Philadelphia.

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

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

Bisani, M. and H. Ney. 2004. Bootstrap estimates for confidence intervals in asr performance evaluation. In *Proceedings of ICASSP'04*, volume 1, pages 409–412, May.

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

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

Core, M. G. and J. F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In David Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. AAAI.

Dybkjær, L. and W. Minker, editors. 2008. *Recent Trends in Discourse and Dialogue*, volume 39 of *Text, Speech and Language Technology*. Springer, Dordrecht.

Fraser, M. and G. Gilbert. 1991. Simulating speech systems. *Comp. Speech Lang.*, 5:81–99.

Fukada, T., D. Koll, A. Waibel, and K. Tanigaki. 1998. Probabilistic dialogue act extraction for concept based multilingual translation systems. In *Proceedings of International Conference on Spoken Language Processing*, volume 6, pages 2771–2774.

Godfrey, J., E. Holliman, and J. McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proc. ICASSP-92*, pages 517–520.

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

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

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

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

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

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

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

Webb, N. and Y. Wilks. 2005. Error analysis of dialogue act classification. In *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, pages 451–458.

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

Williams, J. D. and S. Young. 2007. Partially observable markov decision processes for spoken dialog systems. *Comput. Speech Lang.*, 21(2):393–422.

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

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