A Study of a Segmentation Technique for Dialogue Act Assignation*

Carlos-D. Martínez-Hinarejos
Instituto Tecnológico de Informática, Universidad Politécnica de Valencia
Valencia, Spain
cmartine@dsic.upv.es

1 Introduction

A dialogue system is usually defined as a computer system that interacts with a human user to achieve a task using dialogue [5]. In these systems, the computer must know the meaning and intention of the user input, in order to give the appropriate answer. The user turns must be interpreted by the system, taking only into account the essential information, i.e, their semantics for the dialogue process and the task to be accomplished. This information is usually represented by labels called Dialogue Acts (DA) [2] which label different segments of the turn known as utterances [8]. The DA labels usually take into account the semantics of the utterance with respect to the dialogue process, but they can include semantic information related to the task the dialogue is about.

Therefore, the correct assignation of DA to a user turn is crucial to the correct behaviour of the dialogue system. Several models have been proposed to perform this assignation. In the recent years, probabilistic models have gained importance in this task [8]. In the assignation task, these models are applied on non-annotated dialogues. Most of the previous work done on the assignation of DA is performed on user turns segmented into utterances, although the availability of the segmentation is not usual in the dialogue corpora nor in a real dialogue system.

The proposed assignation models can be easily adapted to the lack of segmentation into utterances. In this article, we present a model based on

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

Proceedings of the 8th International Conference on Computational Semantics, pages 299–304,
Tilburg, January 2009.

Hidden Markov Model (HMM) and N-grams that can be applied to segmented and unsegmented turns. The results show that the lack of segmentation causes many errors in the assignation of DA. Therefore, we propose another model based on N-grams (NGT model) that segments a user turn into utterances previously to the DA assignation.

2 Models

The statistical model that provides the DA assignation uses the current turn word sequence W and the previous DA sequence U' to obtain the optimum DA sequence $\widehat{U}$ that maximises the posterior probability of U, i.e., $\widehat{U} = \text{argmax}_U \Pr(U|W, U')$. This formula can be developed as presented in [6] to obtain:

$$\widehat{U} = \underset{U}{\text{argmax}} \sum_{r, s_1^r} \prod_{k=1}^{r} \Pr(u_k | u_{k-n-1}^{k-1}) \Pr(W_{s_{k-1}+1}^{s_k} | u_k) \tag{1}$$

In this equation, $\Pr(u_k|u_{k-n-1}^{k-1})$ can be modelled as an N-gram (of degree n) and $\Pr(W_{s_{k-1}+1}^{s_k}|u_k)$ as a HMM. This model can be used when there is an available segmentation by simply eliminating the sum and prod operators and fixing the s_k values to those provided by the segmentation.

The NGT model [7] can provide a segmentation of a dialogue turn. This model is inspired in the GIATI methodology [3], which relies on the concept of alignment between the input and the output sentences. In our case, the input sequence is the dialogue turn, and the output sequence is the utterances boundaries. A re-labelling process is applied on this pair of sequences to convert them to an unique sequence where input and output are joined.

From this sequence that includes the utterance boundaries, an N-gram can be inferred. We implemented the Viterbi search to work directly on the N-gram that acts as a transducer, and gives the name to the model: N-gram transducers (NGT).

Therefore, given an NGT model and an unsegmented turn, a segmentation of the turn into utterances can be obtained by using the Viterbi search. In the next section we present the experiments that verify the accuracy of the segmentations provided by this model along with the results the HMM-based model provides with these segmentations.

3 Experiments and discussion

In this section, we compare the performance of the HMM-based model when using the correct segmentation into utterances, the NGT segmentation and when no segmentation of the turns is given. The experiments were performed on the SwitchBoard corpus [4] and on the Dihana corpus [1] (with the two and three-level labels). All the experiments were performed using a cross-validation approach. Some simplifications were performed on the corpora before training the models and performing the experiments.

The HMM model is trained for each DA label from the utterances of the training dialogues annotated with the corresponding label. The N-gram model is trained from the sequences of DA labels in the training dialogues. In the NGT model, the training dialogues were re-labelled to obtain the sequences of words with the attached utterance boundaries. From these sequences, different N-grams were inferred with N values from 2 to 5. These N-grams were used to obtain the segmentation of the turns in the test dialogues into utterances.

The decoding step was performed using the Viterbi algorithm. In the case of the segmented dialogues, the segmentation was fixed to that provided in the manual annotation or by NGT. In the case of the unsegmented turns, the search was performed on the complete turn and the optimal DA assignation provided the segmentation as a by-product.

We can compute the accuracy of the segmentation given by the NGT model and the HMM-based model from the reference segmentation. These results are presented in Table 1, along with the error in the number of utterances (for bigram). In general, the NGT model provides a more accurate segmentation of the dialogue turns. The reduction of the segmentation errors is quite significant, with relative reductions of more than a 20%. This is in general true for the estimation of the correct number of utterances (except for Dihana 2-levels). With these results, we can expect that the decoding of the HMM-based model using the NGT segmentation would be of higher quality than the decodings on the unsegmented turns.

The results are shown in Table 2. The evaluation is done with two measures: complete turn DA error (all the labels must be coincident) and DAER (like WER for speech recognition systems but at the DA level). As was expected, the number of erroneously labelled turns increases with the unsegmented approach for both measures. This relative increase is lower in the simpler dialogue corpora than in the more complex corpora. The results for the NGT model are with a bigram, and they show that in the case of the SwitchBoard corpus, the quality of the DA assignation is quite better,

Table 1: Segmentation errors for complete turn with the NGT and HMM-based model, and errors in the number of estimated utterances for bigrams. Best results for each corpus are shown in boldface.

Model	Corpus / Ngram dgr.	2	3	4	5	Utt. error
NGT	SwitchBoard	**26.1**	26.8	28.9	31.6	23.0
	Dihana 2-levels	**9.6**	10.0	10.1	9.7	9.6
	Dihana 3-levels	13.6	11.9	12.4	**11.3**	12.8
HMM	SwitchBoard	**41.4**	41.6	41.8	42.1	34.7
	Dihana 2-levels	14.2	13.9	13.9	**13.8**	6.0
	Dihana 3-levels	38.8	38.6	38.6	**38.6**	17.3

but in the rest of cases the improvement is not significant or even negative (Dihana 2-level).

The results show that the clearest source of errors in DA assignation that the NGT model can produce is due to the split of the turn into a wrong number of utterances. This clearly induces an erroneous assignment of the DA sequence to the turn, as the number of labels will be different to the reference and, consequently, the turn will be counted as erroneous. Even in the case of DAER error, this fact is present: for Dihana 2-levels, the unsegmented labelling provides a lower error rate than the NGT-segmented one. This is sound with the error in the number of segments that provide the HMM-based and the NGT model (Table 1, last column).

Therefore, we can see a clear correlation between the assignation of an incorrect number of utterances and the error rates in DA assignation when using NGT-segmented turns or unsegmented turns. Consequently, we can conclude that segmenting into the correct number of utterances is critical to obtain a correct assignation of DA. The DAER rate presented in Table 2 confirms that considering only the turns with an incorrect segmentation but a correct number of utterances, many of them present a correct assignation of DA.

The main conclusion we can extract from these experiments and results is that a correct hypothesis on the number of utterances of a dialogue turn is needed to obtain a correct assignation of DA to that turn, and that the accuracy of the segmentation is not so important.

Following these conclusions, future work will be directed to the obtainment of models that, given a dialogue turn, provide its number of utterances.

Table 2: Errors for complete turn DA assignation and DAER results (for bigrams) with different segmentation conditions. Best results for each corpus are shown in boldface.

Segmentation	Corpus / Ngram dgr.	2	3	4	5	DAER
Correct	SwitchBoard	38.3	**38.1**	38.7	39.7	38.3
	Dihana 2-levels	21.6	21.1	**20.8**	21.0	20.1
	Dihana 3-levels	24.6	**24.0**	24.2	24.8	24.7
NGT	SwitchBoard	48.1	**48.0**	48.5	49.3	43.2
	Dihana 2-levels	27.1	26.8	26.6	**26.4**	29.0
	Dihana 3-levels	30.6	**30.0**	30.1	30.7	30.6
Unsegmented	SwitchBoard	59.6	**59.0**	59.6	61.0	60.7
	Dihana 2-levels	24.6	**24.3**	24.4	24.6	24.0
	Dihana 3-levels	32.5	**31.6**	31.9	32.4	33.2

This hypothesis on the number of utterances can be used to restrict the exploration of possible segmentations in both the presented models and it can help to obtain better results.

Other work can be done to improve the models or the use of other segmentation models. This can help us to obtain more general conclusions on the importance of segmentation for the DA assignation task.

References

[1] N. Alcacer, J.M. Benedí, F.Blat, R.Granell, C.D. Martínez, and F.Torres. Acquisition and Labelling of a Spontaneous Speech Dialogue Corpus. In *SPECOM*, pages 583–586, Patras,Grecia, 2005.

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

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

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

[5] J. Van Kuppevelt and R. W. Smith. *Current and New Directions in Discourse and Dialogue*, volume 22 of *Text, Speech and Language Tech.* Springer, 2003.

[6] C-D. Martinez-Hinarejos, J-M. Benedi, and R. Granell. Statistical framework for a spanish spoken dialogue corpus. *Speech Comm.*, 50:992–1008, 2008.

[7] C.D. Martínez-Hinarejos. Automatic annotation of dialogues using n-grams. In *Proceedings of TSD 2006*, LNCS/LNAI 4188, pages 653–660, Brno, Czech Republic, Sep 2006. Springer-Verlag.

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