Evaluation of the incremental dialogue annotation using N-gram Transducers

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

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

Abstract

The annotation of dialogues in terms of Dialogue Acts (DA) is an important task in the development of dialogue systems. Recently, the N-gram Transducers (NGT) technique showed a better performance than other techniques in the annotation of unsegmented dialogue transcriptions. However, this technique has not been employed in an incremental fashion, which is closer to the annotation framework. In this work, we checked the performance of NGT in this incremental framework and the influence of the size of the partitions in the effort of annotating the whole SwitchBoard corpus.

1. Introduction

One interesting application of natural language processing is dialogue systems [1]. A dialogue system is a computer system that interacts with a human user to fulfil a task whose completion requires several interactions. The behaviour of the dialogue system is defined by the dialogue strategy, which defines the reactions of the system to the user input. The user input is generally interpreted in terms of Dialogue Acts (DA) [2], which are labels that define the intention and the involved data in a subsequence of the input (usually known as segment). DA can be extended to system interactions, to reflect the actions that the system carries out.

The dialogue strategy can be based on statistical models [3], whose parameters are estimated from dialogues annotated with DA. This statistical approach is more flexible than a classic rule-based approach, but requires a large amount of annotated data to accurately estimate the models. Consequently, the annotation of a training dialogue corpus is one of the biggest efforts in the construction of a dialogue system. In the last decade, some works presented statistical models [4, 5] to speed-up this annotation process: the automatic annotation models are used to obtain a draft annotation that is corrected by the human annotator, which supposes a lower effort than annotating the dialogue from scratch. One of the most powerful annotation techniques is the NGT (N-gram Transducers) model [7], which uses an N-gram derived from the joining of words and DA and another N-gram derived from sequences of DA to obtain the DA annotation of unlabelled dialogue turns.

However, experiments reported in those previous works use a large training set and a small test set [4, 6]. In this work we present results using the NGT model in an incremental fashion, i.e., a small set of dialogues is used to train the models, another set is annotated with these models, the annotation is corrected and the corrected dialogues are added to the training set for the next step. This process is closer to the usual annotation framework. The results will show that, although there is a degradation in performance, the NGT model is still a reasonable tool to speed-up the complete annotation of a dialogue corpus. Moreover, the results demonstrate that a small amount of data annotated from scratch is more convenient, in terms of effort, than using a larger amount.

This paper is organised as follows: Section 2 provides an overview of the NGT model; Section 3 describes the corpus used for the experiments (SwitchBoard); Section 4 defines the experimental framework and shows the results of the experiments; Section 5 provides some conclusions and reveals possible work lines.

2. The NGT dialogue annotation model

The N-gram Transducers (NGT) model [7] is based on the inference of an N-gram from a set of extended symbol sequences. These extended symbols are build from the alignment of the symbols of a parallel corpus of input-output sequences. In the case of a dialogue corpus, the input symbols are the words and the output symbols are the DA labels, which are usually aligned to the last word of the segment they label. From the extended sequences an N-gram can be inferred. This N-gram can be used to process an unlabelled input sequence (sequence of words) and associate the corresponding DA labels to each possible segment. The Viterbi decoding process is shown in Figure 1.

In this labelling process each word is taken to build its corresponding tree level. Each node is expanded into o nodes, where o is the number of different outputs (DA labels) that were associated to the word in the training samples (including the empty output). The probability of each node is recalculated according to the probability of the parent node, the probability given by the N-gram of extended symbols and the probability of the associated sequence of DA labels (given by another N-gram model of DA sequences, see [7] for a detailed description of the probability computation). The output of the decoding process is the sequence of words with the corresponding attached DA labels. This provides an annotation of the input sequence (dialogue) along with its segmentation. The NGT implementation used in this work is publicly available in [8].

3. The SwitchBoard corpus

The SwitchBoard corpus [9] is a corpus of human-to-human conversations by telephone in English. It includes spontaneous speech conversations about general topics, without a clear task to complete, with frequent interruptions, background noises, hesitations and non-linguistic sounds (such like laughter). The final corpus consists of 1,155 dialogues, with approximately 115,000 turns, and a vocabulary size about 42,000 words.

This corpus was manually transcribed (including special annotation for the previously described phenomena) and annotated at the dialogue level using the SWBD-DAMSL scheme [10], a simplified version of the standard DAMSL (Di-

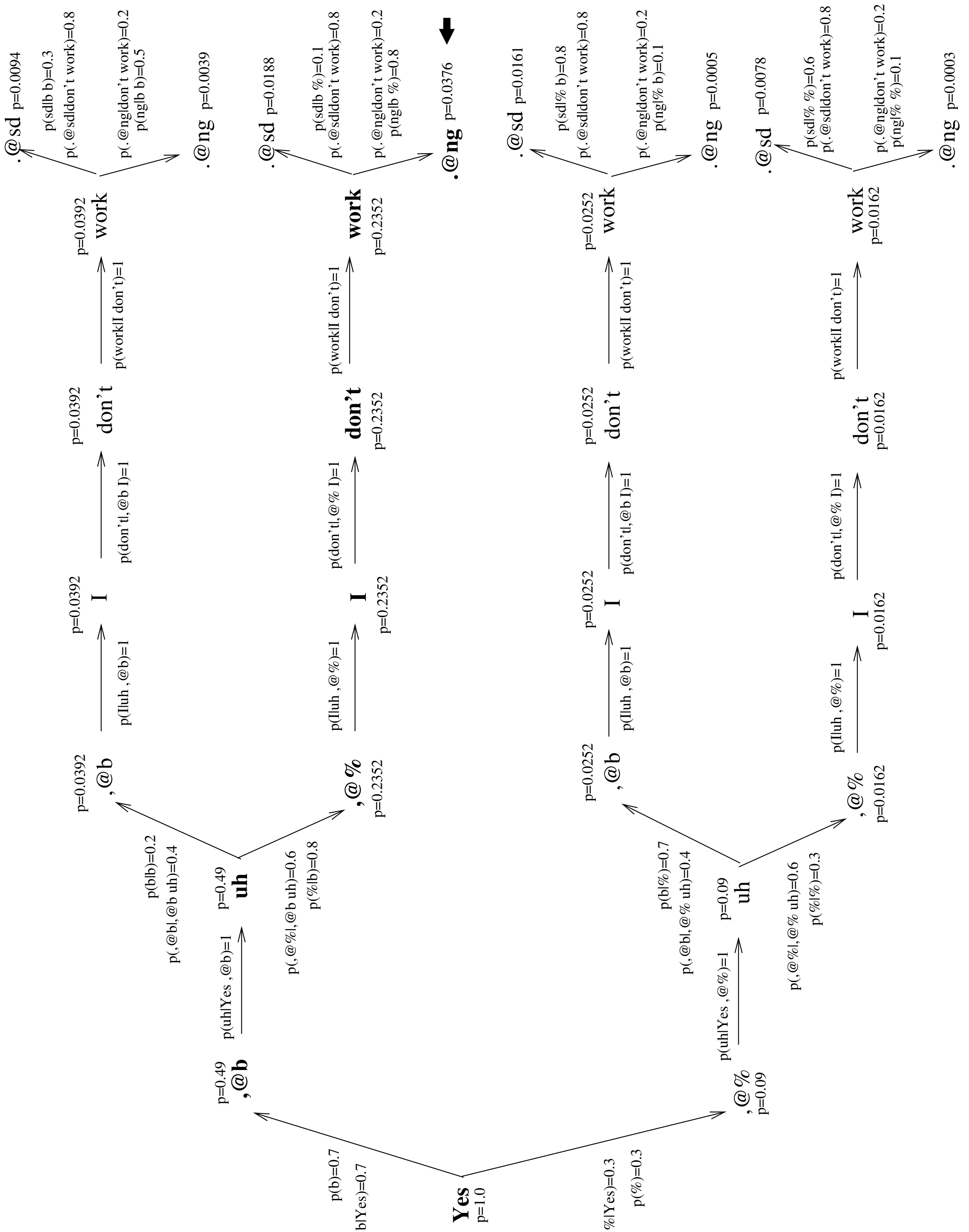

Figure 1: An example of the Viterbi tree search for the NGT model for the sentence "Yes, uh, I don't work.". Symbols before @ are words and symbols after @ are DA labels (in this case, *b*-backchannel, %-uninterpretable, *sd*-statement-non-opinion, and *ng*-negative-non-no-answer). Best hypothesis is in boldface and marked by the dark arrow. In this example, trigrams are used in all models.

alogue Act Mark-up in Several Layers) annotation set [11]). SWBD-DAMSL includes 42 different DA labels that represent several communicative functions, such as statement, question, backchannels, etc., and subcategories of these functions (e.g., statement opinion/non-opinion).

4. Experiments and results

The objective of our experiments is to verify the appropriateness of the NGT model for the annotation of the SwitchBoard corpus in an incremental fashion. This analysis is useful to check whether NGT is convenient to be used in an actual annotation task and to adapt the technique to an active learning [14] (the selection of the most informative samples for training) or interactive-predictive framework [13] (the use of information given by the user to obtain a better search in the model).

The annotation task usually starts from a set of transcribed dialogues on which the human annotators must place the DA labels according to a set of predefined rules. To use a statistical annotation model to obtain draft annotations, an initial set of dialogues must be annotated from scratch. The parameters of the statistical model are inferred from this initial set, and the model is applied to a new set of unlabelled dialogues. These dialogues are revised by the human annotators to correct the possible errors. The correctly annotated dialogues are added to the previously annotated set, and this new set is used to improve the estimation of the parameters of the models. This cycle is repeated until the entire set of dialogues is correctly annotated.

Consequently, the annotation framework employs an incremental training set, whose size is initially much smaller than the complete corpus and becomes larger in each cycle. This is in contrast with the approach taken by many previous works [4, 7], where training sets are usually composed of a large number of dialogues from the corpus.

In our experiments we used this incremental approach to verify the appropriateness of the NGT model. We initially compared the incremental approach with the standard cross-validation approach. To simplify the framework, we used incremental partitions of regular size on the SwitchBoard corpus. The complexity of the transcription of the SwitchBoard corpus was lowered down by removing the interruptions and overlaps (and joining the corresponding interrupted turns), transcribing the words to lowercase and separating punctuation marks.

The partitions were based on those used in the cross-validation approach. The first comparison we made was in terms of annotation error rates. In the annotation task both the correct label and the correct position are important. Consequently, we adopted the SegDAER (Segmentation and Dialogue Act Error Rate) measure. In this measure, the sequences to be compared are formed by the DA labels joined with their position in the turn. SegDAER is the average edit distance between the correct sequence and the sequence obtained by the annotation model, and was used in previous works on the measure of the quality of annotation errors [7].

In this case, we compared the results for each partition using the cross-validation approach and the incremental approach, using 11 partitions of 105 dialogues each one. This allows us to avoid the possible differences given by the specific difficulty of each partition. In the cross-validation approach, each training set is formed by 10 partitions, whereas in the incremental approach, the training set for annotating the n-th partition in the sequence is composed of the $n-1$ previous partitions.

Following the best results reported by the cross-validation approach, the experiment was performed using trigrams as NGT model and DA language model. The results, along with the 90% confidence intervals (obtained by bootstrapping with 1000 repetitions, following the method in [12]), are presented in the graph in Figure 2. The first partition is not included in the incremental approach as its SegDAER is 100%.

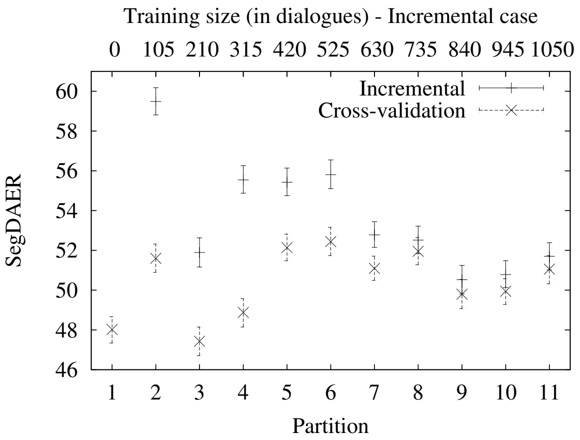

Figure 2: SegDAER comparison between the cross-validation approach and the incremental approach using 11 partitions.

From these results we can see that the absolute difference in SegDAER in each partition is lower than the 8% in all cases (and lower than 5% in most partitions). These differences become insignificant from the partition 8 (735 dialogues in the training set). Consequently, we can conclude that the NGT technique presents a moderate degradation (which disappears when using approximately 60% of the whole data) in performance when using the incremental approach, and that it is still useful in the dialogue annotation framework.

Another interesting experiment is related to the effect on the global annotation effort dependence on the partition size. In this case, we introduce a global annotation effort measure based on the SegDAER of each partition: the Relative Histogram Error Area (RHEA). This measure is based on measuring the percent of the area of the error histogram for each partition and dividing it by the area of the worst-case error histogram (all partitions present a SegDAER of 100%), i.e., RHEA=$\frac{\sum_i E_i S_i}{S_T}$, where E_i is the SegDAER of partition i, S_i is the size of partition i (the size of the individual partition, not the cummulative sum of the sizes of the previously annotated partitions) and S_T is the size of the whole corpus. The lower bound for this measure is defined by the ratio between the size of the partition annotated from scratch and the total size of the corpus (RHEA=$\frac{100S_1}{S_T}$). The size of the partitions and the corpus can be measured in different terms, such as number of dialogues, number of turns and number of words, among others. In any case, RHEA reductions can be considered as proportional reductions in the number of errors that must be corrected by human annotators.

We computed the RHEA measure for five different sets of partitions (of 3, 5, 7, 11, and 21 partitions each set). We computed RHEA using dialogue and turn as the basic size unit for each partition. The results using 3-grams and 4-grams for each model are presented in Table 1.

From these results, a clear conclusion is that the smaller the size of the partition, the lower the annotation effort. This is congruent with the intuitive idea, as the main effort is related to the

Table 1: RHEA results for the different sets of partitions using 3 and 4-grams as NGT and DA N-grams. The size of the partitions is defined in terms of number of dialogues and number of turns.

NGT N-gram		3				4			
DA N-gram		3		4		3		4	
Num. part.	Dial/Part	Dial	Turn	Dial	Turn	Dial	Turn	Dial	Turn
3	385	69.4	69.5	69.3	69.5	71.0	71.1	71.0	71.1
5	231	62.5	63.0	63.0	63.5	64.9	65.4	64.9	65.4
7	165	60.2	61.0	60.4	61.2	62.0	62.8	62.2	62.9
11	105	57.9	58.5	58.1	58.7	60.2	60.8	59.9	60.5
21	55	56.4	56.8	56.3	56.6	58.4	58.8	58.0	58.3

annotation from scratch, and the correction of a draft annotation requires, in general, less effort. From the results we can see that the optimal combination of N-grams for the NGT and the DA N-gram depends on the size of the partitions, but differences are really small for the same NGT N-gram degree. We can see that measuring the size of the partitions in dialogues or turns does not change the optimal combination of models and the conclusions on the best partition size.

5. Conclusions and future work

In this work we presented the use of a dialogue annotation technique (NGT) in a more realistic incremental framework, in order to compare its behaviour with respect to that in the cross-validation approach. Results showed that, although there is a statistically significant increment of annotation error, it is not a dramatical increment that disregards the use of the annotation technique. We evaluated the effort of the annotation from the SegDAER of each partition, and results demonstrate that the smaller the size of the partitions, the lower the effort.

Future work is directed to studying these elements in other corpora and applying the technique in an interactive-predictive framework to make a more user-oriented evaluation [13]. Another direction is related to the use of active learning [14] to obtain an appropriate selection of the partitions to be annotated from scratch and the annotation order of the rest of the dialogues, in order to reduce the annotation effort even more.

6. Acknowledgements

This work was partially supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV “Consolider Ingenio 2010” program (CSD2007-00018), the MITTRAL (TIN2009-14633-C03-01) project, and the FPI BES-2007-16834 grant. Also supported by the EC (FEDER) and the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grants Prometeo/2009/014, GV/2010/067, and ACOMP/2010/051.

7. References

[1] L. Dybkjær and W. Minker, Eds., *Recent Trends in Discourse and Dialogue*, ser. Text, Speech and Language Technology. Dordrecht: Springer, 2008, vol. 39.

[2] H. Bunt, “Context and dialogue control,” *THINK Quarterly*, vol. 3, 1994.

[3] J. D. Williams and S. Young, “Partially observable markov decision processes for spoken dialog systems,” *Comput. Speech Lang.*, vol. 21, no. 2, pp. 393–422, 2007.

[4] 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.

[5] N. Webb and Y. Wilks, “Error analysis of dialogue act classification,” in *Proceedings of the 8th International Conference on Text, Speech and Dialogue*, 2005, pp. 451–458.

[6] 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.

[7] C.-D. Martínez-Hinarejos, V. Tamarit, and J.-M. Benedí, “Improving unsegmented dialogue turns annotation with n-gram transducers,” in *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC23)*, vol. 1. Hong Kong: City University of Hong Kong Press, Dec. 2009, pp. 345–354.

[8] C.-D. Martínez-Hinarejos, “The ngt dialogue annotation software,” 2010, http://users.dsic.upv.es/˜cmartine/research/resources/ngt.tgz.

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

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

[11] 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*, D. Traum, Ed. Menlo Park, California: AAAI, 1997, pp. 28–35.

[12] M. Bisani and H. Ney, “Bootstrap estimates for confidence intervals in asr performance evaluation,” in *Proceedings of ICASSP’04*, vol. 1, May 2004, pp. 409–412.

[13] E. Macklovitch, “Transtype2: The last word,” in *Proceedings of the 5th International Conference on Languages Resources and Evaluation (LREC 06)*, Genoa, 2006, p. 167172.

[14] A. Beygelzimer, S. Dasgupta, and J. Langford, “Importance weighted active learning,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, Montreal, Quebec, Canada, 2009, pp. 49–56.