Active Learning for Dialogue Act Labelling

Fabrizio Ghigi[1], Vicent Tamarit[2], Carlos-D. Martínez-Hinarejos[2] and José-Miguel Benedí[2]

[1] Dpto Electricidad y Electrónica, Facultad de Ciancia y Tecnología, Universidad del País Vasco, Sarriena s/n, 48940, Leioa, Spain, fabrizio.ghigi@gmail.com

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

Abstract. Active learning is a useful technique that allows for a considerably reduction of the amount of data we need to manually label in order to reach a good performance of a statistical model. In order to apply active learning to a particular task we need to previously define an effective selection criteria, that picks out the most informative samples at each iteration of active learning process. This is still an open problem that we are going to face in this work, in the task of dialogue annotation at dialogue act level. We present two different criteria, weighted number of hypothesis and entropy, that we have applied to the Sample Selection Algorithm for the task of dialogue act labelling, that retrieved appreciably improvements in our experimental approach.

1 Introduction

Dialogue systems are an important application in the field of Natural Language Processing. A dialogue system is usually defined as a computer system that interacts with a human by using dialogue to achieve a defined objective [Dybkjær and Minker, 2008]. The computer system interprets the user input in the form of dialogue meaningful units, which are usually known as Dialogue Acts (DA) [Bunt, 1994], and that are used by the system to determine its reaction to user input (this reaction can be coded in DA labels as well). The reaction of the system is defined by the dialogue strategy, which indicates what actions the system must perform, including the response generation to the user. These strategies can be rule-based strategies [Gorin et al., 1997] (based on a set of predefined rules) or data-based strategies [Young, 2000] (based on statistical models). In any case, these strategies are based on the study of dialogues of the task to be fulfiled, and in their annotation in terms of DA. The goal of this work is to explore various sample selection criteria and employ them in an active learning strategy framework for dialogue annotation. The results prove that we can achieve a good annotation model performance by using only

```
Yes , from Madrid at 10:30 .
    ↑              ↑       ↑
(Acc:Dep-t) (Ans:Org) (Ans:Dep-t)
------------------------------------------------------------
Yes ,@(Acc:Dep-t) from Madrid@(Ans:Org) at 10:30 .@(Ans:Dep-t)
```

Fig. 1. An alignment between a dialogue turn and its corresponding DA labels (from the DIHANA task) and the result of the re-labelling process, where @ is the attaching metasymbol.

a subset of the initial set of samples (the most effective data samples), reducing the effort needed to label the dialogues that will be used to train the final dialogue model. The automatic annotation method used in this work is the N-Gram Transducer (NGT) annotation model, described in [Tamarit et al., 2009]. We report experiments to find a good selection criterion for the Active Learning Algorithm [Hwa, 2000] for the task of automated DA labelling of the DIHANA corpus [Benedí et al., 2006].

This document is organised as follows: In Section 2, the statistical model for labelling the unsegmented dialogue turns is presented. In Section 3, active learning strategy is introduced. In Section 4, the selection criteria choosen are presented. In Section 5, the experimental setting used to test the learning criteria and the obtained results are detailed. In Section 6, final conclusions and future work are presented.

2 The NGT annotation model

The dialogue annotation problem can be presented as, given a word sequence $\mathcal{W}$ that represents a dialogue, obtain the sequence of DA $\mathcal{U}$ that maximises the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$. This probability can be modelled by a Hidden Markov Model approach by using the Bayes rule [Stolcke et al., 2000] or by directly modeling the posterior probability $\Pr(\mathcal{U}|\mathcal{W})$.

The NGT model 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[3] [Casacuberta et al., 2005]. 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).

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

In the case of dialogues, the input language is the sequence of words of the dialogue, the output language is the sequence of DA of the dialogue, and the alignment is between the last word of the segment and the corresponding DA. Thus, for each turn $w_1 w_2 \ldots w_l$ and its associated DA sequence $u_1 u_2 \ldots u_r$, the re-labelling step attaches the DA label to the last word of the segment using a metasymbol (@), providing the extended word sequence $e_1 e_2 \ldots e_l$, 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 1 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, the alignments between the words in the turn and the corresponding DA labels are monotonic (no cross-inverted alignments are possible). Consequently, no conversion to SFST is necessary to efficiently apply a search algorithm on the n-gram, since for each input word we can decide whether to emit or not a DA label without referring to posterior words. Therefore, this n-gram acts as a transducer and gives the name to the technique (NGT: N-Gram Transducers) [Martínez-Hinarejos et al., 2009].

The decoding in the NGT model is a Viterbi search which forms a search tree. The i-th level of the tree corresponds to the i-th input word in the sequence. Each input word is expanded for all the possible outputs it has associated in the alignments in the training corpus. The probability of each branch is updated according to the corresponding parent node, the n-gram probability of the corresponding extended word sequence and the n-gram probability of the corresponding DA sequence (in case a new DA is produced).

In the final step, the search on the NGT model produces a search tree where each leaf node represents a possible solution (an annotation hypothesis) to the annotation problem for the input word sequence (a dialogue). Each leaf node has associated a probability calculated by the method described above, and the leaf node with highest probability is taken as the optimal solution for the annotation problem. The solution is obtained by going up from the leaf node till the root node of the constructed tree, giving an annotation and a segmentation on the dialogue.

3 Active Learning

Active learning selects more data at each iteration of the learning process from the unlabeled set by asking someone to manually label that data. The algorithm stops when no more data or no more human resources are available, or a sufficient performance is reached.

In order to apply the active learning algorithm, a criterion that allows our system to assign a "priority" to each sample in the unlabeled set data is needed; then we can use the given scores to sort the set of unlabeled data, and choose a subset with higher priority (according to the selected criterion). The selected samples are manually labelled and they are used to reestimate the model parameters. If the accuracy goal is not overtaken, the reestimated model is used in the next step of sample selection. Otherwise, the process is finished.

In our implementation of Active Learning Algorithm [Hwa, 2000], U is a set of unlabeled candidates; L is a small set of labeled training samples; M is the current model.

```
Initialize
    $M \leftarrow Train(L)$
Repeat
    $N \leftarrow Select(n,\ U,\ M,\ f)$
    $U \leftarrow U\text{-}N$
    $L = L \cup Label(N)$
    $M = Train(L)$
Until $(M=M_{true})$ or $(U=\varnothing)$ or (Human Stops)
```

4 Sample Selection Criteria

In our case, the training process of the model is the usual training for the NGT model, and the labelling process in the human annotation of the dialogues. Consequently the key point of the Active Learning Algorithm presented in Section 3 is the sample selection criterion. Depending on the task, various criteria could be used. In this work we tested the algorithm with two different criteria: Weighted Number of Hypothesis and entropy. Both criteria are based on the idea that the more significant samples that we can add to the training set are those samples that are more difficult to assign a correct label. These "difficult" samples can be measured by the "uncertainty" in finding a correct label for the sample.

4.1 Weighted Number of Hypothesis

The first criterion is the number of hypothesis retrieved by the NGT decoding. Each hypothesis gets a weight, that depends on the feasibility of the hypothesis: the most probable hypothesis have more weight on the final decision, while the less probable hypothesis not strongly affect our uncertainty. We use for each sample the following equation:

$$\sum_i \frac{\Pr_i(x)}{\Pr_{max}(x)} \tag{1}$$

where $\Pr_i$ represents probability of i-th hypothesis obtained by the decoding of sample x (in our case a possible decodification of the current dialogue in DA) with the current model, and $\Pr_{max}(x)$ is the maximum probability among all hypothesis of the current sample. When this value is computed for each unlabeled sample, we select the dialogues with highest scores of "uncertainty". We decide to assign this value to each hypothesis because not every hypothesis retrieved by the model adds the same uncertainty: hypothesis with higher probability get a weight close to 1, while hypothesis less probable get less weight.

4.2 Entropy

The second criterion used is that of *Entropy*. The *Entropy* is a common way in language processing of evaluating language models. It measures how difficult is for the model to recognize a specific sample: the smaller the entropy, the easier for the model to decode correctly the sample. The entropy for a dialogue is computed according to the following expression [Robinson, 2008]:

$$H_m(t) = -\frac{1}{\Pr_m(s)} \left(\sum_{t \in T} \Pr_m(t) \log \Pr_m(t) \right) + \log \Pr_m(s) \qquad (2)$$

where $\Pr_m(s)$ is the word sequence probability by the model M(in our case, given by a n-gram of words), $\Pr_m(t)$ is the probability of the decodification retrieved by the model M(i.e., the probability given by the NGT model), and T is the dialogue set.
To have homogenous values, the computed value of *Entropy* is normalized by the lenght (number of words) of the current sample, because the entropy value is influenced by the length of the sample. Like previous criterion, *Entropy* gives us an indication of how much we know about the current sample, i.e., an uncertainty level.

5 Experiments

Experiments are developed for the dialogue act annotation task.The automatic annotation method used in this work is the NGT model.The learning criteria described in Section 4 are tested on the Dihana corpus that will be described in Section 5.1. In order to evaluate results we use DAER and SegDAER metrics. DAER is the average edit distance between the reference DA sequences of the turns and the DA sequences assigned by the labelling model. SegDAER is an average edit distance between sequences derived from the reference and the annotation result; in this case, sequences are a combination of the DA label and its position (segmentation). Incremental selection of training samples is lead by the Active Learning Algorithm described in Section 3.

5.1 Dihana Corpus

The Dihana corpus [Benedí et al., 2006] is a set of spoken dialogues in Spanish language, between a human and a simulated machine, acquired with the Wizard of Oz (WoZ) technique. It is restricted at the semantic level (dialogues are related to the task of obtaining information about train tickets), but natural language is allowed (there are no lexical or syntactical restrictions). The Dihana corpus is composed of 900 dialogues about a telephone train information system. It was acquired from 225 different speakers (153 male and 72 females), with small dialectal variants. There are 6,280 user turns and 9,133 system turns. The vocabulary size is 823 words. The total amount of speech signal is about five and a half hours. The annotation scheme used in the corpus is based on the Interchange Format (IF) defined in the C-STAR project [Lavie et al., 1997], which was adapted to dialogue annotation. Details on the annotation process are available in [Alcácer et al., 2005].

5.2 Experiment Strategy

We have used the same partition (720 dialogue to pick up for training, 180 dialogues for test) of the Dihana corpus for every experiment developed, maintaining the following strategy:

1. Start the experiment with a small training set, picked out by a general criteria (in fact we picked out the two largest dialogues).
2. Train a model with this small training set, and verify the accuracy of the system, calculating DAER and SegDAER for the NGT model predictions.
3. With current model compute score for each remaining dialogue in unlabeled set, using criteria *Weighted Number of Hypothesis*, equation (1), or *Entropy*, equation (2).
4. Select a subset of the remaining dialogues with higher scores.
5. Include the selected dialogues in the training samples.
6. Return to step 2.

The Sample Selection Algorithm described in Section 3 is used to manage incremental selection of training samples.

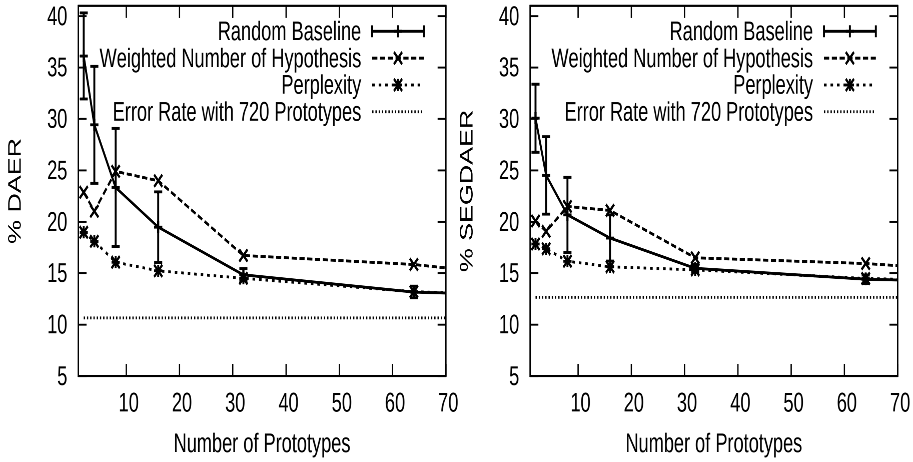

Fig. 2. Comparison among Random Baseline, *Weighted Number of Hypothesis*, equation (1), and *Entropy*, equation (2), error rate (DAER and SegDAER) behaviour while incrementing the training set size. The lowest line represent the error rate using the entire set of availables dialogues (in our case 720).

5.3 Results

In Figure 2 we can see DAER and SegDAER trends for Random Baseline, and for the two selection criteria, *Weighted Number of Hypothesis*, equation (1), and *Entropy*, equation (2).

As we can clearly see from Figure 2, *Entropy* criterion is a very effective selection criterion for this task, it has better performance than Random Baseline until the asymptote is reached. Moreover, performance is close to that obtained with the whole training set. We tried more experiments incrementing training set size by one dialogue at each iteration, but this does not change significantly the error rate trend.

6 Conclusions and Future Work

In this document we have shown results of applying Active Learning to a dialogue act labelling task. We have seen that choosing a well founded criterion (*Entropy*) to implement Active Learning Algorithm, significant performance boost can be achieved. In the experiments developed *Entropy* criterion obtained really good results, while *Weighted Number of Hypothesis* criterion had a variable behaviour, although more experiments should be perform in the future to confirm its properties.

Future work contemplates the application of presented criteria against other corpora (such as SwitchBoard) to confirm goodness of criteria, the parallelization of the Active Learning Algorithm to speed up the selection process, the exploration of other selection criteria, the application of this work in an interactive framework and the analysis of the error rate for each single dialogue act label taking into account its frecuency in the corpus.

Acknowledgments. Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), MITTRAL (TIN2009-14633-C03-01) projects and the FPI scholarship (BES-2009-028965). Also supported by the Generalitat Valenciana under grant Prometeo/2009/014 and GV/2010/067.

References

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

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

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

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

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

Gorin et al., 1997. Gorin, A., Riccardi, G., and Wright, J. (1997). How may I help you? *Speech Comm.*, 23:113–127.

Hwa, 2000. Hwa, R. (2000). Sample selection for statistical grammar induction. In *Proceedings of the 2000 Joint SIGDAT*, pages 45–52, Morristown, NJ, USA. Association for Computational Linguistics.

Lavie et al., 1997. Lavie, A., Levin, L., Zhan, P., Taboada, M., Gates, D., Lapata, M. M., Clark, C., Broadhead, M., and Waibel, A. (1997). Expanding the domain of a multi-lingual speech-to-speech translation system. In *Proceedings of the Workshop on Spoken Language Translation, ACL/EACL-97*.

Martínez-Hinarejos et al., 2009. Martínez-Hinarejos, C. D., Tamarit, V., and Benedí, J. M. (2009). Improving unsegmented dialogue turns annotation with N-gram transducers. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC23)*, volume 1, pages 345–354.

Robinson, 2008. Robinson, D. W. (2008). Entropy and uncertainty. 10:493–506.

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

Tamarit et al., 2009. Tamarit, V., Benedí, J., and Martínez-Hinarejos, C. (2009). Estimating the number of segments for improving dialogue act labelling. In *Proceedings of the First International Workshop of Spoken Dialog Systems Technology*.

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