Active Learning to Speed-up the Training Process for Dialogue Act Labelling

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

[1]Dpto Electricidad y Electrónica, Facultad de Ciencia y Tecnología,
Universidad del País Vasco, Sarriena s/n, 48940, Leioa, Spain
[2]Instituto Tecnológico de Informática, Universitat Politècnica de València,
Camino de Vera s/n, 46022, Valencia, Spain
fabrizio.ghigi@gmail.com,{cmartine,jbenedi}@iti.upv.es

Abstract

The dialogue act labelling task is the process of splitting and annotating a dialogue into dialogue meaningful units; the labelling task can be performed semi-automatically by using statistical models trained from previously annotated dialogues. The appropriate selection of training dialogues can make the process faster, and Active Learning is one suitable strategy for this selection. In this work, Active Learning based on two different criteria (Weighted Number of Hypothesis and Entropy) has been tested for the task of dialogue act labelling by using the N-gram Transducers models. The framework was tested against two heterogeneous corpora, DIHANA and SwitchBoard. The results confirm the goodness of this kind of selection strategy.
Index Terms: dialogue act labelling, active learning, data uncertainty

1. Introduction

A spoken dialogue system is a conversational agent able to have a talk with a human, with the perspective to achieve a predetermined goal. While performing the setup process of a dialogue system an indispensable condition for the success of a data-based strategy (Young, 2000) is the availability of a big amount of annotated dialogues. Annotating a dialogue corpus in terms of Dialogue Acts (DA) (Bunt, 1994) is one of the most expensive, time-consuming and annoying tasks while developing a dialogue system. The common scenario is a situation where an abundant amount of unlabeled data is available, and labelling this data is an expensive task in terms of human effort and time. An alternative to this manual annotation is provided by the use of semi-automatic annotation tools which provide a draft annotation that must only be revised by the human annotators. These annotation tools (most of them based on statistical models such as those described in (Stolcke et al., 2000)) can speed-up the annotation process and consequently the construction of a whole dialogue system. To develop an automatic DA Labelling system, we need training data, i.e., dialogues segmented in terms of DAs, permitting to perform the learning process. Usually, the more training data we have, the better performance the system can reach.

In such scenario, it could be desirable to have a criterion that permit us to select just the most informative samples to be manually labeled, reducing the amount of data we need to label in order to reach a good performance for our annotating system. The main idea is to manually label the set of dialogues that will provide a better statistical annotation model, but having a compromise between the amount of human-tagged data and the overall accuracy of an automatic tagging system. Therefore, a criterion must be formulated to obtain the most informative dialogues with respect to a given statistical annotation model. This criterion can be iteratively applied to the remaining unlabeled samples, to select the samples that according to the current model can produce larger improvements in the system performance generating a new model, until a target performance has not been reached.

In general, this problem can be formulated as: having a set of samples we want to use to train a classifier, and consequently need to be manually tagged, how to reduce the cost of the tagging process. Time and human resources needed for the tagging process, as costs, are proportionals to the number of samples we want to tag. Research on a selection criterion for sample selection is still an open problem, especially in dialogue, and probably task-dependent.

Therefore, we are looking for an effective criterion that permits to select just the most informative and significant samples for the task we are approaching, DA labelling. In such a scenario the application of the *Active Learning* technique (Hwa, 2000) could be useful in order to reduce the labelling task costs.

In this work we are going to present the results of applying *Active Learning* for the task of DA labelling by using the N-gram Transducers models. Two different uncertainty based criteria, Weighted Number of Hypothesis and Entropy, are tested and compared against a random baseline to check their appropriateness in the Active Learning sample selection. Results are obtained on the transcriptions of two different spoken dialogue corpora, DIHANA (Benedí et al., 2006) (human-computer, semantically restricted) and SwitchBoard (Godfrey et al., 1992) (human-human, semantically not restricted).

2. Dialogue Act Labelling with the N-Gram Transducers Model

DA Labelling is the task of segmenting a dialogue into dialogue meaningful units (segments) and associating to each segment a label (DA) depending on the dialogue-related meaning of that segment.

The DA Labelling 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 $\Pr(\mathcal{U}|\mathcal{W})$, for example by using the N-Gram Transducers (NGT) model (Martínez-Hinarejos et al., 2009).

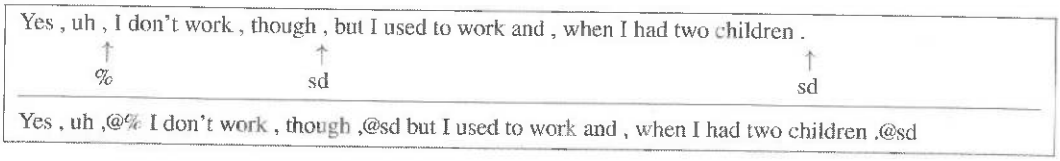

Figure 1: An alignment between a dialogue turn and its corresponding DA labels (from the SWBD-DAMSL scheme, %: uninterpretable, sd: statement-non-opinion), and the result of the re-labelling process, where @ is the attaching metasymbol.

The NGT model estimates $\Pr(\mathcal{U}|\mathcal{W})$ 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 (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 which combine the words in the input and output sentences. The corpus of extended words is used to infer a grammatical model (usually a smoothed n-gram).

When dealing with dialogues, the input and output language are formed by the words and the DA labels of the dialogue, respectively. Each DA label is aligned to the last word of its corresponding segment. Thus, for each turn $w_1 w_2 \dots w_l$ and its associated DA sequence $u_1 u_2 \dots u_r$, the re-labelling step attaches the DA label to the last word of the segment using a metasymbol (@). The result of the process is 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, $e_i = w_i @ u_k$ 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 (usually, a smoothed n-gram) and converted into a SFST. In the case of dialogues, since alignments between words and DA labels are monotonic (no cross-inverted alignments are possible), 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 or not to emit a DA label without referring to posterior words). Therefore, this n-gram acts as a transducer and gives the name to the technique.

A Viterbi search decoding is employed on the NGT model to obtain the dialogue annotation. This decoding process builds a search tree, in whose i-th level is represented 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, giving as many branches as possible outputs. 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).

At the end of the search process, a full search tree is produced. In this search tree each leaf node represents a possible solution (an annotation hypothesis) to the annotation problem for the input dialogue. Each leaf node has associated a probability, 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 tree, giving an annotation and a segmentation on the dialogue.

3. Active Learning

Using machine learning algorithms we are able to develop systems that can increase their performance by adding more training data. According to this, we can use our system in an active way, choosing the most appropriate actions to improve the model performance in the fastest and cheapest possible way. In case of active learning (Riccardi and Tür, 2003), we iteratively improve system performance by adding new training data the system can learn from. The system results on unlabelled data can be used to select the samples that would provide a more effective parameter estimation for the model, i.e., suggest which samples must be annotated and added to the training set.

In the description of Active Learning algorithm (Hwa, 2000) below, U is a set of unlabeled candidates, L is a small set of labeled training samples, M is the current model, M_{true} is a model that achieve an objective performance in the labelling task, f represents the selection criteria chosen, n the number of unlabeled samples selected at this iteration, N the new selected set we are going to label:

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

4. Sample Selection Criteria

The key point in the active learning algorithm is the selection criteria. In this section two selection criteria are briefly described: (1) Weighted Number of Hypothesis, and (2) Entropy. The two criteria try to estimate the uncertainty of the sample, i.e., how difficult is for the current model to recognise the sample. In any case, the samples with highest uncertainty will be those to be selected.

4.1. Weighted Number of Hypotheses

This criterion utilizes the set of different hypotheses retrieved by the NGT decoding to compute a score for each dialogue remaining in the unlabeled set. For each dialogue, each hypothesis (leaf node of the search tree) is normalised by the maximum probability among all the hypotheses for dialogue (the most likely hypothesis). This causes that less probable hypotheses have a lower impact on the final computed score (since hypotheses with small probability do not affect much to uncertainty). The uncertainty is computed using the following expression for each dialogue in the unlabeled set:

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

In this equation, $\Pr_i$ represents the probability of i-th hypothesis (a possible decodification of the current dialogue in DA using the current model), and $\Pr_{max}(x)$ is the probability of the most likely hypothesis for this sample x. After calculating this

uncertainty for each remaining unlabeled sample, the subset of n dialogues with the highest uncertainty is selected for the next labelling step.

4.2. Entropy

Entropy gives a measure of how difficult finds the system to recognise a specific sample. It is used in several natural language processing tasks to evaluate language models. A lower value of entropy reflects the facility for the system to decode the sample. In our case, the expression that was used to compute the entropy values is the following (Robinson, 2008):

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

with $\Pr_m(s)$ the n-gram probability according to the model M, $\Pr_m(t)$ the NGT decoding probability and T the whole set of hypothesis retrieved by the NGT model. Since this value depends on the length of the dialogue, it is normalised by the length of the current sample. After computing the entropy value for each unlabeled dialogue, those dialogues with highest entropy (uncertainty) values are chosen for the next labelling step.

5. Experiments

Experiments are developed using two heterogeneous corpora, DIHANA (Benedí et al., 2006) and SwitchBoard (Godfrey et al., 1992), that permits us to confirm the goodness of the selection strategy; Active Learning is performed for both criteria, and results (Section 5.4) are compared against a Random Baseline obtained by calculating average and variance of six random experiments with different seeds; the metric chosen to evaluate system performance is SEGDAER, described in Section 5.3.

5.1. Corpora

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).

SwitchBoard Corpus. The SwitchBoard corpus (Godfrey et al., 1992) is a set of spoken dialogues in English Language, human-human conversations by telephone not related to a specific task; it includes 1,155 different conversations, performed by 500 different speakers. The number of turns in the dialogues is around 115,000; in average, each turn has 1.8 segments. The vocabulary size is approximately 42,000 words. It was annotated using a shallow version of the DAMSL (Core and Allen, 1997) annotation scheme, 42 different labels present in the SWBD-DAMSL (Jurafsky et al., 1997). These labels represent categories such as statement, backchannel, questions, answers, etc.

5.2. Strategy

For both corpora, DIHANA (Benedí et al., 2006) and SwitchBoard (Godfrey et al., 1992), we have performed Active Learning; for DIHANA we have used 180 dialogues as test and 720 dialogues as training, for SwitchBoard 105 dialogues as test and 1050 for training. The strategy implemented that perform Active Learning (Section 3), follows these steps (U is the input set of unlabeled samples, L is the labelled samples set, and M the draft model):

1. Train an initial model M from a small set of tagged samples (set L), picked out by a general criteria (in fact we picked out the two dialogues with more turns).
2. Compute SEGDAER (see subsection 5.3) for the NGT model predictions of the system, according to the current model.
3. Apply a function f over the unlabeled set of samples that, according to M and to the selection criteria chosen (Weighted Number of Hypothesis (1) or Entropy (2)), computes a score for each dialogue remaining in the unlabeled set U
4. Select a subset N of these dialogues with the highest scores.
5. Take out the set N from the unlabeled set U.
6. Manually label the set N (in this case this step is simulated, no human resources were employed, the entire labeled set was available).
7. Add the labeled set N to the labeled set L.
8. Reestimate the model M with the new set L.
9. If sufficient performance is not reached and there are still unlabeled samples and human resources, restart from the first step of loop. In this case, we stop the algorithm when no more samples are available.

We have chosen to use an exponential function to determine size of new samples set to select, in fact 2^i, where i is the index of current iteration. This incremental size of selection is desirable because of the asymptotic behavior of the error rate; thus, with this incremental size approach we can see the improvements with small amounts of training data, checking how fast we can converge to the asymptote, while adding more data to a large training set does not strongly affect the error rate. The Sample Selection Algorithm described in Section 3 is used to manage incremental selection of training samples.

5.3. Evaluation Metrics

To evaluate the system performance we use the SEGDAER metric: it is the average edit distance between the reference DA sequences of the turns and the DA sequences assigned by the labelling model; in this case, sequences are a combination of the DA label and its position, which means that it takes into account also the dialogue segmentation, because is important in DA labelling task not only to predict the correct labels, but also to put them in the correct position in the dialogue.

5.4. Results Analysis

5.4.1. SEGDAER

This section presents the results obtained, using Active Learning for DA Labelling task, against the two corpora considered, DIHANA (Benedí et al., 2006) and SwitchBoard (Godfrey et al., 1992), reporting in each graphic the SEGDAER behavior obtained with the two criteria described in Section 4, compared against the Random Baseline; the lowest horizontal line in the graphics shows the error rate obtained training the system with the entire training set available. Random baseline allows to compare the effect of making an appropriate selection instead of not using any criteria for the selection of samples (an upper bound); bottom line allows to compare the effect of selecting against using (and annotating) all the training data (lower bound).

In Fig. 2 is represented the error rate, computed in terms of SEGDAER (Subsection 5.3), obtained in each iteration, so the higher the score, the worse the performance. As we can clearly see in the results in Fig. 2, the error behavior is asymptotic, fact confirmed by the small variance of random experiments after 5 iterations. This means that we can reach a good performance in the earlier steps of Active Learning process, and the system performance remain almost unaffected when adding more training data after a small number of iterations of the Active Learning algorithm (Section 3). The results shows that Entropy (2) selection criterion works very well in this task, performing better than Random Baseline in both corpora tested, while the Weighted Number of Hypothesis (1) criterion had a variable behavior, retrieving performance similar to Entropy using SwitchBoard corpus, and worse than random behavior testing with DIHANA corpus.

5.4.2. Performance

With Active Learning we aim to speed up the labelling process, labelling just the most informative samples, but maintaining the system performance; the focus is on save up money and time in the labelling process, but using just a subset of training sample we do not expect improvements in the error rate, we just expect to achieve a performance as close as possible to the performance obtained training with the entire training set available.

In these graphics we present the results obtained in the experiments in another perspective, showing the percentage of training set we need to achieve a given percentage of the final performance obtained training the system with the entire set available.

The "performance" of the system is just the inverse of the SEGDAER, e.g. $100 - SEGDAER$, and then normalized to a percentage, taking the final SEGDAER obtained training with the entire training set available as the 100% of the performance as shown in (3).

$$\frac{(100 - SEGDAER) * 100}{(100 - FINALSEGDAER)} \quad (3)$$

6. Conclusions

In conclusion we applied Active Learning to DA labelling task, for two very heterogeneous corpora, DIHANA and SwitchBoard, using uncertainty based criteria to perform samples selection at each iteration of the Active Learning algorithm. Results obtained in the two different domains confirm the goodness of the uncertainty based criteria.

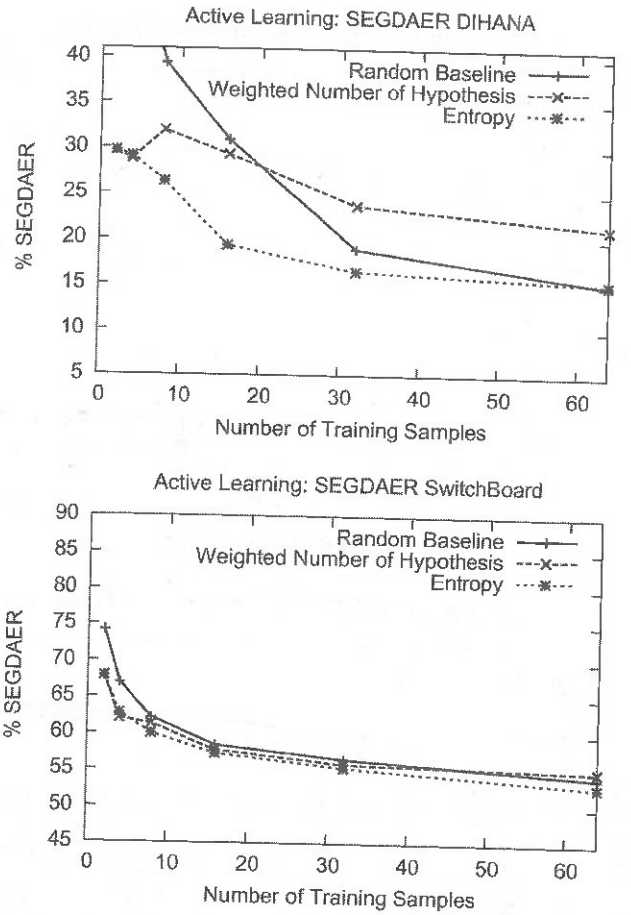

Figure 2: *SEGDAER results while performing Active Learning for DIHANA and SwitchBoard. Graphics include results for the two criteria tested, Weighted Number of Hypothesis and Entropy, compared against the Random Baseline, and a lower bound that represents the error rate obtained by training the model with the entire set available. The results obtained converge asymptotically to the lower bound, and about 64 and 256 prototypes the results present no significant differences against using all the training data in DIHANA and SwitchBoard corpora, respectively. It is worth emphasizing the significant improvement obtained for the corpus DIHANA using Active Learning implemented with Entropy criterion in the first iterations up to 60 training samples.*

For both corpora used the experimental results shows that we can achieve a good performance, 95% of the performance obtained training with the entire set available, labelling just the 20% of the samples in the unlabeled set. According to these results we can save an important amount of time and money in the labelling task by using Active Learning with uncertainty based criteria to select the most informative samples.

This kind of approach is reusable in other statistical models where we can compute the scores for the two criteria proposed to implement the Active Learning algorithm, the only elements needed to compute these scores are the set of hypotheses with their probabilities, and also an n-gram model in the case of Entropy criterion.

Planned future work includes parallelization of Active Learning algorithm, exploration of other selection criteria, ap-

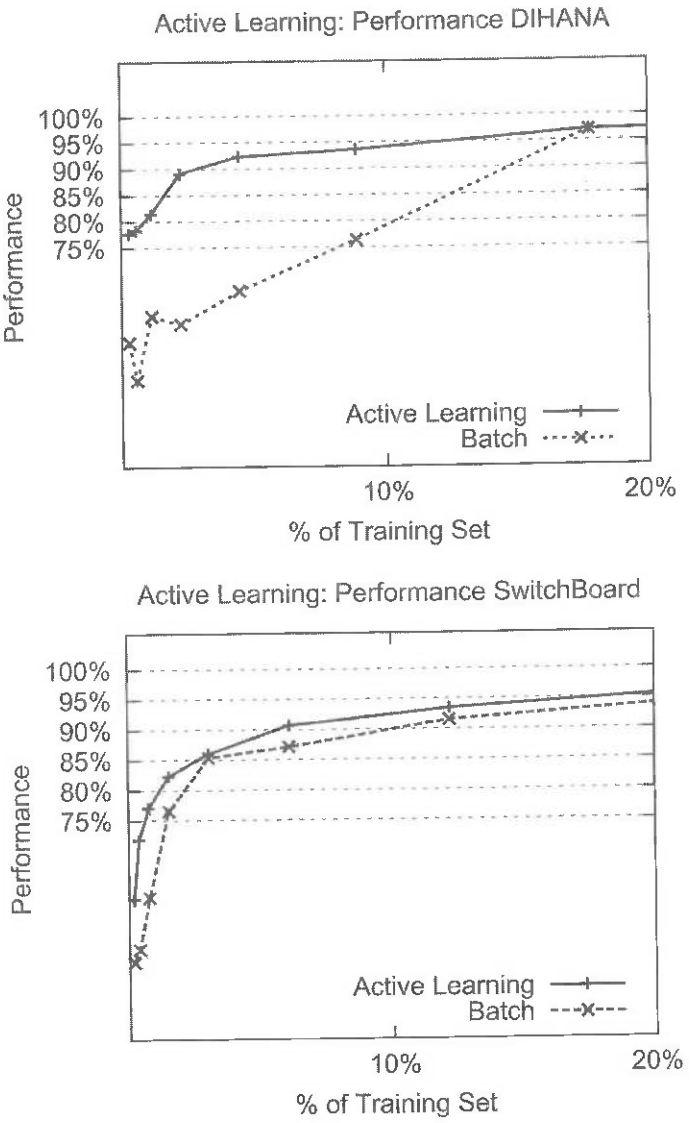

Figure 3: *The Performance of the system, computed at each iteration of the Active Learning algorithm, shows that we can achieve 95% of the final performance by using just the 20% of the entire training set available, for both corpora used, DIHANA and SwitchBoard. In the figures is represented the performance obtained at each iteration of the Active Learning algorithm implemented with Entropy criterion and the performance obtained with a Batch training process. In this graphic are shown just the results obtained implementing the Active Learning algorithm with Entropy criterion because this is the criterion that has obtained best performance in the experiments, allowing us to better appreciate the improvements achieved with Active Learning technique. Has to be noticed the significant performance improvements obtained in the first iterations implementing the Active Learning algorithm with the Entropy criterion in the DIHANA corpus.*

plication in an interactive framework, and the analysis of single DA label error rate.

7. Acknowledgements

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, the FPI scholarship (BES-2009-028965) and the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project. Also supported by the Generalitat Valenciana under grant Prometeo/2009/014 and GV/2010/067.

8. References

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, 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, H. (1994). Context and dialogue control. *THINK Quarterly*, 3:19–31.

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

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

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

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.

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

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*, pages 67–72.

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.

Riccardi, G. and Tür, D. (2003). Active and unsupervised learning for automatic speech recognition. In *INTERSPEECH*.

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

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.

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