Active Learning for Dialogue Act Classification

Björn Gambäck[1,2]*, Fredrik Olsson*[1]*, Oscar Täckström*[1]

[1]SICS, Swedish Institute of Computer Science AB, Kista, Sweden
[2]Computer and Information Science, Norwegian University of Science and Technology
{gamback,fredriko,oscar}@sics.se

Abstract

Active learning techniques were employed for classification of dialogue acts over two dialogue corpora, the English human-human Switchboard corpus and the Spanish human-machine Dihana corpus. It is shown clearly that active learning improves on a baseline obtained through a passive learning approach to tagging the same data sets. An error reduction of 7% was obtained on Switchboard, while a factor 5 reduction in the amount of labeled data needed for classification was achieved on Dihana.

The passive Support Vector Machine learner used as baseline in itself significantly improves the state of the art in dialogue act classification on both corpora. On Switchboard it gives a 31% error reduction compared to the previously best reported result.

Index Terms: dialogue acts, active learning, SVMs

1. Introduction

The paper describes a study on applying active learning techniques to the task of automatically labelling dialogue acts. A *dialogue act* is a semantic unit expressing the communicative intention of a dialogue participant (human or machine). Correctly classifying the dialogue turns can improve the performance of dialogue managers, speech recognisers, and other systems. A range of machine learning approaches have thus been used for the task; mainly supervised techniques having in common that they at random select a subset of the unannotated data to query.

In contrast, the present paper investigates the use of *active learning* for dialogue act tagging. Active learning is also a supervised machine learning technique, but one in which the learner is in control of the data subset used for learning. That control is utilized to ask an *oracle* (e.g., a human with extensive knowledge of the domain at hand) about the labels of the instances for which the model learned so far makes the most unreliable predictions. The active learning process takes as input a set of labeled examples, as well as a larger set of unlabeled examples, and produces a classifier and a relatively small set of newly labeled data. The overall goal is to create as good a classifier as possible, without having to annotate more data than necessary.

The next section introduces the active learning technique. Then Section 3 gives a brief overview of the dialogue corpora employed for the experiments. Section 4 first describes the experimental setup, then shows results from experiments using active learning and compare them to results reached using a regular (passive) learning strategy. Finally, in Section 5 conclusions are drawn and possible future directions suggested.

2. Active Learning

The aim of active learning is to keep the manual annotation effort at a minimum. Essentially, it is about making use of what is already known in order to find out what is new and informative. The learner should only ask the oracle for advice where the training utility of the result of such a query is high. Active learning is an iterative process which starts with a small set of labeled data and a large set of unlabeled data. In each iteration, a base learner configuration is used to train a classifier on all the labeled data available. The classifier is then applied to the unlabeled data, and the most informative instances in that data are selected and handed over to the oracle for manual annotation. The manually labeled instances are added to the training data and the process starts over again, with a slightly larger set of training data. At some point, the active learning is terminated and a final classifier is obtained by applying the base learner configuration to all training data available at that point.

The active learning paradigm can utilize a range of machine learning strategies for the base learner. When it is necessary to distinguish between "ordinary" machine learning and active learning, the former is sometimes referred to as *passive learning* or learning by *random sampling* from the available set of labeled training data. An active learner is considered superior to its passive counterpart if it generates a learning curve which is steeper and dominates (lays above) the learning curve generated by the passive learner. Theoretically, if the data set is separable, an active learner would be able to reduce the distance between the guess and the true threshold at an exponential rate, while a passive learner only would be able to do it at a linear rate [1]. In practical applications, and with non-separable data, things are not as clear, though. It has even been claimed that active learning is not beneficial at all in dialogue act classification [2]. However, that hypothesis is clearly falsified by the results below.

The overall properties of active learning have been addressed by several authors, e.g., [3] and [1]. For a literature survey of active learning in general see [4], while [5] gives an overview of its application to language processing. The utility of active learning for speech processing has been shown by several researchers, e.g., within the EC/FP6 spoken language understanding project LUNA (IST 33549) where an active learner only needed 30% of the annotated data to reach the same performance level as a passive learner trained on the entire data set [6]. Actual experiments with active learning helping humans annotate the LUNA corpus with dialogue act tags using a Conditional Random Fields-based classifier showed a factor 3 speed-up in the average time to annotate a dialogue with the support of the classifier [7].

3. Data Sets

In the present experiments to validate the use of active learning for dialogue act tagging, two annotated dialogue corpora were used, the English human-human telephone speech Switchboard corpus and the Spanish human-machine Dihana corpus. Both these corpora have previously been used for training dialogue act classifiers, but with passive, supervised learning approaches.

3.1. The Switchboard corpus

Switchboard [8] is a large multispeaker corpus of conversational speech over telephone. The conversations are over a wide range of topics and domains. The corpus contains some 2,500 conversations by 500 speakers of American English, with a set of longer dialogues with more than 1 hour each of speech from 50 of the speakers, and many shorter dialogues of about 5 minutes each from the others. The recordings are accompanied by a time-aligned word-by-word transcription. The experiments in this paper used a subset of the transcribed corpus containing 224,000 utterances from 1,155 of the shorter conversations tagged with dialogue acts from the SWBD-DAMSL tagset [9]; a subset of the corpus also used in many previous dialogue act tagging experiments (e.g., [10] and [11]). The original SWBD-DAMSL tagset contains 226 classes, but these can be clustered into 44 dialogue act types [9] . The most-frequent label assignment baseline then is 36% (for the category <STATEMENT-NON-OPINION>).

The label set used (and whether some labels are merged or ignored) makes it difficult to compare dialogue act tagging approaches applied to a specific corpus. Previous experiments on Switchboard have tended to reduce the number of classes by merging and/or filtering, commonly to 42 classes after merging the two tags '%' and '%-' (<UNINTERPRETABLE> and <ABANDONED>), and filtering out the '+' () tag [10]. The segment class makes up about 8% of the corpus and is problematic since it indicates overlapping speech. The tagging accuracy for it can be as low as 47.66% [11]. Two common approaches to treating this have been to remove all utterances tagged '+' or to ignore the tag by merging an utterance tagged '+' with the preceding utterance by the same speaker; in both cases reducing the corpus size (to about 205,000 utterances). In contrast, all 44 classes of dialogue acts in the clustered SWBD-DAMSL tagset were used in the experiments reported in this paper.

3.2. The Dihana corpus

The Dihana corpus [13] consists of spontaneous speech human-computer (wizard-of-Oz) telephone dialogues in Spanish about train timetable information and fares. The corpus consists of 900 dialogues with 225 users and about 5.5 hours of user speech, with a vocabulary of 823 words (48,243 words in total). There are 9,179 user and 13,829 wizard utterances which have all been manually annotated with dialogue acts. The dialogue act annotation scheme used for Dihana is based on three levels. Level 1 corresponds to the speaker's intention (speech act), Level 2 represents the implicit information referred to in Level 1, and Level 3 contains the specific data provided in the utterance. Using these levels and distinguishing between user and system labels, there are 248 different labels (153 for the user, 95 for the system). Combining only Level 1 and 2, there are 72 labels (45 user, 27 system), and 16 labels (7 user, 9 system) with only Level 1.

4. Active Learning Experiments

Dialogue act classification using active learning was evaluated over the Switchboard and Dihana corpora, and compared to passive learning for the same task. The classifiers were implemented in MALLET, a freely available Java library of a range of machine learning methods.[1] MALLET was used for feature extraction and as framework for running experiments, while the actual training utilized the open source LIBLINEAR package[2] for which a java wrapper was written. LIBLINEAR is a C++ library for large-scale linear classification which supports logistic regression and linear Support Vector Machines (SVMs). All learning experiments utilized an L_2-regularized linear SVM.

[1] `http://mallet.cs.umass.edu`

[2] `http://www.csie.ntu.edu.tw/~cjlin/liblinear`

4.1. Training data encoding and selection

Previous approaches to dialogue act tagging have used a range of features, most commonly lexical ones deduced from the words of the utterance to classify, and information about the dialogue context (the preceding and/or succeeding utterances with corresponding dialogue acts). For the present experiments, the following features were used to represent each instance: Word uni-, bi-, and trigrams of the instance; utterance initial / final word uni-, bi-, tri- and 4-grams; presence of *wh*-words ('what' | 'why' | 'where' | 'which' | 'who' | 'how'), exclamation marks ('!'), or question marks ('?'). For Switchboard, additional features were extracted from the transcription to encode the presence of `<info>`, `<<info>>`, `+`, `--`, `{ ... }`, and `[ ... ]` transcription codes. (Note that it might be difficult to extract these latter features automatically in a real system.) Context was modeled by adding the full set of features of the previous utterances, concatenating each feature with the utterance offset, to the current instance. Since the dialogue acts following the learning instance are not taken into account, the classification can be done on-the-fly, so that the classifier would be applicable to run a dialogue manager in a real-time application setting.

In order to ensure annotation consistency and to allow for faster experiments, the human oracle was simulated in the active learning experiments; the unlabeled data was actively selected from a set of pre-labeled data instead of manually inspected by a human annotator on-the-fly. Note that this is to the *disadvantage* of the active learning model, since in a larger set of unlabeled data there can be more informative instances that might be ignored by selecting from a pre-labeled data set. The active learning set-ups utilized a $[-1, 0]$ context-window only, including in the learning process the feature representation of the current instance and of the instance immediately preceding it. A single (rather than committee-based) learner was trained on batches of data: at each iteration a small set of annotated data instances was added to the training set available to the learner. The minimal-distance-to-hyperplane heuristic [14] was used to select the instances to be annotated by the oracle, namely the instances closest to decision boundaries (i.e., those balancing between two or more classes).

4.2. Active Learning on Switchboard

Figure 1 shows the results of using active learning on Switchboard, for the first 20,000 instances (about 10% of the actual data set of 210,000 dialogue acts). The horizontal line at the top of the figure indicates the maximum accuracy which can be obtained using passive learning on the entire corpus. The other two curves show how the dialogue act tagging results improve as more annotated training data is added, in the case of active (upper, dashed curve) and passive learning (lower, filled curve). The graphs are averages of the values obtained after ten runs over the data set for each learning strategy (10-fold cross validation).

The *batch size* in the experiments shown in the graph was set to 100, that is, at each iteration the training data set available to the learner was augmented by 100 samples. Thus the active learning curve in itself consists of 200 learning experiments making up the samples at different learning points. The amount of data available to the learner at the start of the first iteration, the *seed size*, was also set to 100.

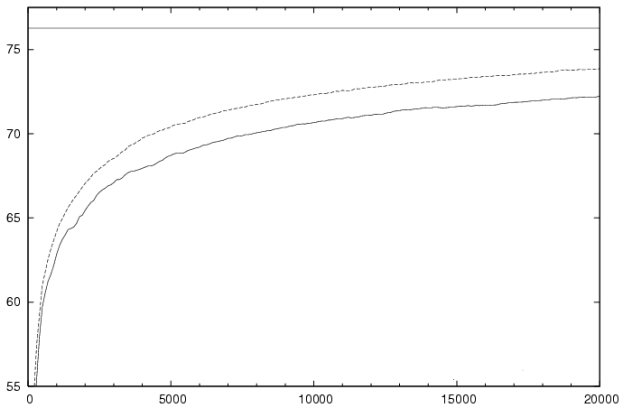

Figure 1: Switchboard learning curves (accuracy vs number of instances in the training set); maximum obtainable by passive learning on the entire corpus (horizontal line), active learning (upper curve), and passive learning (lower curve).

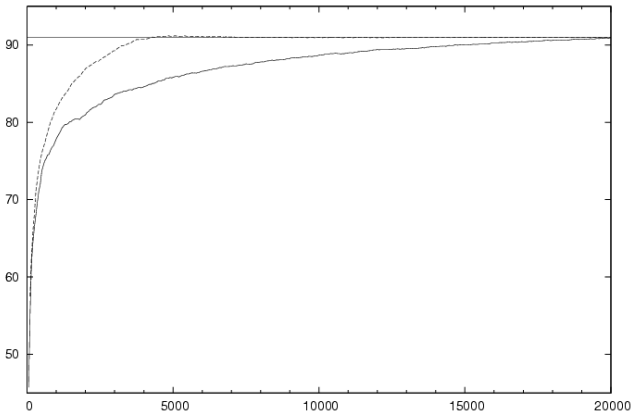

Figure 2: Dihana learning curves; maximum (horizontal line), active learning (upper curve), and passive learning (lower).

As can be seen in Figure 1, the active learning curve dominates the passive curve throughout; fairly constantly with a difference of about 2% of accuracy, equivalent to an error reduction of around 7% for most of the portion of the curves seen in the figure (upward from about 4,000 training instances). In contrast to previous experiments on Switchboard, all 44 dialogue act classes (i.e., without class merging and/or filtering) have been used in the experiments reported here.

Notably, the passive learning baseline used here for Switchboard with just a $[-1, 0]$ context-window (bigram), is above the state-of-the-art: using only the representation of the previous utterance as context it reaches a value in these experiments of 76.26% (with a standard deviation of 0.27) on the full 44-class tagset. This dialogue act tagging accuracy can be improved to 77.85% (± 0.26) by adding one more previous utterance to the context (thus looking at trigrams). The previously best reported compatible dialogue act classification results on Switchboard was 65.68%, using a decision tree classifier (J4.8, an implementation of the C4.5 algorithm) on a 43-class tagset, including the ('+') class, but merging the <UNINTERPRETABLE> and <ABANDONED> classes [11]. With the equivalent setting, our SVM classifier returned an accuracy of 76.34% (± 0.25) on the $[-1, 0]$ context thus achieving an error reduction of 31.06%.

By removing all the '+' utterances, Verbree *et al.* (2006) increased the overall accuracy to 70.26% [11]. For this 42-class case the SVM classifier gave a 76.50% (± 0.17) average accuracy, with a $[-2, 0]$ context. Previously, a 3-gram model has been used to achieve 71.0% accuracy, but also after removing all the problematic '+' utterances [10]. This has later been improved to 80.72% accuracy, but then by merging the statement classes <STATEMENT-OPINION> and <STATEMENT-NON-OPINION> (raising the most-frequent label baseline), and including the '+' utterances but merging them with the previous utterance by the same speaker, ignoring the '+' class itself [12].

4.3. Active Learning on Dihana

Applying the same technique to the Dihana corpus, Figure 2 also shows three graphs: one for passive learning (filled), one for active learning (dashed), and one at the top of the figure indicating the maximum accuracy which can be obtained when applying passive learning to the entire corpus. Since the Dihana corpus is smaller than Switchboard, the graphs here show almost the entire data set (20,000 of the 23,000 utterances). Just as in the Switchboard case, the seed size on the Dihana corpus was set to 100. However, since the Dihana corpus is much smaller than Switchboard, the batch size was set to 50 (rather than 100). The active learning curve thus shows the results of 400 iterations.

Again active learning clearly out-performs the passive case: already after having used about 20% of the data (4,000 instances), the active learner has reached the maximum level obtained by the passive learner on the entire corpus, achieving a factor 5 reduction in the amount of labelled data needed for the classification; improving on a factor 4 reduction quoted previously [15]. There is a slight indication that the active learner peaks at a level a bit higher than the passive case, but then falls back in line with it. This is consistent with previous results that an SVM trained on a well-chosen (by active learning) subset of the data often performs better than one trained on the entire corpus [14].

The fact that Dihana is human-machine dialogue (while Switchboard is human-human) explains why dialogue act classification based on the previous dialogue act is more successful for Dihana. Both since the Dihana dialogues are more deterministic, and since a machine (or a human 'wizard' imitating machine behaviour) generates fairly similar text for the same dialogue act, while humans generate sparser data. In general, dialogue act classification on the Dihana corpus is a much easier task than classification on Switchboard. Recall that the Dihana tags are representing three different levels of information. Classifying the utterances using tags from the first two levels with a context of only the previous utterance, the passive learner baseline reaches a 91.61% accuracy (± 0.60), while extending the context to $[-3, 0]$ increases accuracy to 94.08% (± 0.36). For Level 3 this context gives 90.97% (± 0.60) 10-fold average accuracy. Just like in the Switchboard case, the passive learning baseline for Dihana is highly competitive as compared to the state-of-the-art. The previous best dialogue act tagging results on this corpus used trigram HMMs to reach 93.4% accuracy on Level 2 labels and 89.7% on Level 3 (with 5-fold cross-validation) [16].

4.4. Batch vs instance-based learning

As noted above, the active learning experiments reported in this paper were run over *batches* of data, rather than single instances. That is, rather than just adding a single new instance to the training set available to the learner at each iteration, a small set

(a batch) of annotated data instances was added. For the smaller Dihana corpus the batch size was set to 50, while it was set to 100 for Switchboard. These batch-sizes seemed fairly good for both corpora, even though experiments with larger batch sizes (up to 1000 instances) were also run.

Batch-based learning has the effect of speeding up the training process, since fewer iterations are needed. However, it also can produce sub-optimal classifiers, since it is never entirely clear how a whole batch of new instances shall be optimally selected (in particular without interfering with each other). The peak value of 76.26% for Switchboard is for a batch-based run over the data set; however, an instance-based run over the data indicates that even higher accuracy can be reached with this strategy, peaking at 76.50% (with 10-fold cross-validation).

4.5. Structured vs unstructured learning

The dialogue act classification experiments used Support Vector Machines as the underlying strategy. However, active learning as an approach is in itself fairly independent of the machine learning algorithm used for the learning, and it would be interesting to see whether using structured prediction models would improve the results, in particular on the more deterministic Dihana corpus. A small experiment with a structured perceptron on part of the Dihana corpus indicated that a structured model would give better results on that corpus than those obtained by an unstructured model (95.6% vs. 94.1% mean accuracy and 51.0% vs. 34.0% mean exact match, wrt. 10-fold cross validation). However, similar experiments on Switchboard showed no clear benefit. This is in line with previous work on comparing regular and structural SVM on two corpora (Loqui and Enron) which has showed that the structural SVM can give some improvement over the regular SVM, but that this is corpus dependent [17].

5. Conclusions and future work

The paper has evaluated the use of active learning for dialogue act classification over two corpora and compared it to passive learning. The passive learning baseline used here in itself significantly improve the current state of the art in dialogue act classification on Switchboard: the passive Support Vector Machine learner reaches a peak value of 77.85% average accuracy (10-fold cross-validated), while the best reported, compatible result on Switchboard was 65.68% [11]. Still, Figures 1 and 2 show clearly that the active learning case further improves on the passive one. For the Switchboard corpus, the active learner was consistently about 2% better than the passive, giving an error reduction of around 7%; while for the Dihana corpus the active learner only needed about 20% of the available data to reach the maximum level obtained by the passive learner when trained on the entire corpus, thus achieving a factor of 5 reduction in the amount of manually labeled data needed for the classification.

This was achieved even though the hyper-parameters of the learning algorithm were not optimized. A possible candidate for optimisation is the SVM cost parameter, c, which controls the trade off between training errors and model complexity; it creates a soft margin allowing for some classifications errors. The lower the cost of misclassifications, the more flexible and general the model, and faster the training, but with more training errors. Contrastively, larger c values make the classifier more rigid and similar to a hard-margin SVM, and the training slower. Initial experiments both on Switchboard and Dihana indicate that the results can be improved by optimizing it; lower c values do not only lead to faster training, but also to better classification.

6. Acknowledgements

This work was partially carried out within the EC/FP6 integrated project COMPANIONS (IST–34434). Thanks to Ramon Granell (University of Oxford) for providing the Dihana corpus.

7. References

[1] S. Dasgupta, "The two faces of active learning," *Theoretical Computer Science*, vol. 412, no. 19, pp. 1767–1781, Apr. 2011.

[2] A. Venkataraman, Y. Liu, E. Shriberg, and A. Stolcke, "Does active learning help automatic dialog act tagging in meeting data?" in *Proc. 9th Eur. Conf. on Speech Communication and Technology*. Lisbon, Portugal: ISCA, Sep. 2005, pp. 2777–2780.

[3] S. Hanneke, "Theoretical foundations of active learning," PhD Thesis, Machine Learning, Carnegie Mellon University, Pittsburgh, Pennsylvania, May 2009.

[4] B. Settles, "Active learning literature survey," Computer Science, University of Wisconsin, Madison, Tech. Rep. 1648, Jan. 2010.

[5] F. Olsson, "A literature survey of active machine learning in the context of natural language processing," SICS, Stockholm, Sweden, Tech. Rep. T2009:06, Apr. 2009.

[6] P. Gotab, F. Bechet, and G. Damnati, "Active learning for rule-based and corpus-based spoken language understanding models," in *Proc. 11th Workshop on Automatic Speech Recognition & Understanding*. Merano, Italy: IEEE, Dec. 2009, pp. 444–449.

[7] C. Raymond, K. J. Rodriguez, and G. Riccardi, "Active annotation in the LUNA Italian corpus of spontaneous dialogues," in *Proc. 6th Int. Conf. on Language Resources and Evaluation*. Marrakech, Marocco: ELRA, May 2008, pp. 677–678.

[8] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 1. San Francisco, California: IEEE, Mar. 1992, pp. 517–520.

[9] D. Jurafsky, L. Shriberg, and D. Biasca, "Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13," University of Colorado, Boulder, Tech. Rep., Aug. 1997.

[10] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, Sep. 2000.

[11] D. Verbree, R. Rienks, and D. Heylen, "First steps towards the automatic construction of argument-diagrams from real discussions," in *Proc. 2006 Conf. on Computational Models of Argument*. Amsterdam, Holland: IOS Press, Sep. 2006, pp. 183–194.

[12] N. Webb, and M. Ferguson, "Automatic extraction of cue phrases for cross-corpus dialogue act classification," in *Proc. 23rd Int. Conf. on Computational Linguistics*. Beijing, China: ACL, Aug. 2010, pp. 1310–1317. Poster session.

[13] J.-M. Benedí, E. Lleida, A. Varona, M.-J. Castro, I. Galiano, R. Justo, I. López de Letona, and A. Miguel, "Design and acquisition of a telephone spontaneous speech dialogue corpus in Spanish: DIHANA," in *Proc. 5th Int. Conf. on Language Resources and Evaluation*. Genova, Italy: ELRA, May 2006, pp. 1636–1639.

[14] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. 17th Int. Conf. on Machine Learning*, Stanford University, California, Jun. 2000, pp. 839–846.

[15] R. De Mori, F. Bechet, D. Hakkani-Tür, M. McTear, G. Riccardi, and G. Tur, "Spoken language understanding: A survey," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 50–58, Feb. 2008.

[16] C. D. Martínez-Hinarejos, J.-M. Benedí, and R. Granell, "Statistical framework for a Spanish spoken dialogue corpus," *Speech Communication*, vol. 50, no. 11-12, pp. 992–1008, Nov. 2008.

[17] J. Hu, R. J. Passonneau, and O. Rambow, "Contrasting the interaction structure of an email and a telephone corpus: A machine learning approach to annotation of dialogue function units," in *Proc. 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue*. London, England, Sep. 2009, pp. 357–366.