Implicitly-supervised learning in spoken language interfaces: an application to the confidence annotation problem

Dan Bohus[*]
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, 15217
dbohus@cs.cmu.edu

Alexander I. Rudnicky
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA, 15217
air@cs.cmu.edu

Abstract

In this paper we propose the use of a novel learning paradigm in spoken language interfaces – implicitly-supervised learning. The central idea is to extract a supervision signal online, directly from the user, from certain patterns that occur naturally in the conversation. The approach eliminates the need for developer supervision and facilitates online learning and adaptation. As a first step towards better understanding its properties, advantages and limitations, we have applied the proposed approach to the problem of confidence annotation. Experimental results indicate that we can attain performance similar to that of a fully supervised model, without any manual labeling. In effect, the system learns from its own experiences with the users.

1 Introduction

Spoken language interfaces are complex systems that combine many diverse sources of knowledge. Oftentimes, simple algorithmic approaches are insufficient for solving the difficult problems that arise. Instead, machine learning techniques are used, and one of the most often encountered paradigms is that of supervised learning. In this paradigm, the developer provides a training dataset that contains pairs of inputs and desired outputs, and various learning algorithms can be used to derive a model that captures and generalizes the relationship between the two. At runtime, the system generates the corresponding output based on the current input and on the learned model. Such approaches are used in a variety of tasks in spoken dialog systems: acoustic and language-modeling, confidence annotation, dialog act tagging, emotion detection, user modeling, etc.

Supervised learning approaches have however at least two important limitations. First, they require a pre-existing corpus of labeled data. Unfortunately, such corpora are difficult and expensive to collect, especially in the early stages of system development. Secondly, they generally favor an off-line, or "batch" approach. A corpus is collected, manually labeled, and then model parameters are estimated from this data. The resulting model mirrors the properties of the training set, but does not respond well to changes in the system's environment and the underlying data distribution. Unfortunately, such changes are generally expected. Oftentimes, system developers might alter various aspects of system functionality based on feedback and observations. In addition, the users' behavior changes as they repeatedly interact with the system and familiarize themselves with it. Finally, the very introduction of the newly trained model can lead to changes in the interaction. Conversational spoken language interfaces are interactive systems that operate in dynamic environments, and shifts in the underlying data distribution are inevitable.

In this paper, we propose and evaluate a novel learning paradigm that addresses these drawbacks. The proposed approach, dubbed **implicitly-supervised learning**, builds on a key property of spoken dialog systems: their interactivity. The central idea is to extract the required supervision signal from naturally-occurring patterns in the conversation, for instance from user corrections. No developer supervision is therefore required. Rather, the system learns on-line, throughout its lifetime, by interacting with its users. We believe this new para-

[*] Currently at Microsoft Research, Redmond, WA

digm can be applied in a number of learning problems, and can pave the way towards building routinely self-improving systems.

Consider for instance the problem of confidence annotation. Spoken dialog systems use confidence scores to guard against potential misunderstandings: for every utterance, a confidence score reflecting the probability that the system correctly understood the user's utterance is computed. Confidence annotation models are traditionally built using supervised learning techniques (Litman et al, 1999; Carpenter et al, 2001; San-Segundo et al, 2001; Hazen et al, 2002; Hirchberg et al, 2004.) A corpus of dialogs (typically thousands of utterances) is manually labeled by a human annotator: each utterance is marked as either correctly-understood or misunderstood by the system. Supervised learning techniques are then used in conjunction with features that characterize the current utterance to train a model that can predict whether or not this utterance was misunderstood by the system. This approach suffers from the shortcomings we have outlined above: it requires a pre-existing corpus of in-domain utterances, a significant amount of human effort and expertise for labeling this corpus, and it produces a static solution.

The alternative implicitly-supervised solution eliminates these drawbacks. The starting point is the observation that the system could obtain the necessary information (i.e. the misunderstanding labels) by leveraging a particular confirmation pattern that occurs naturally in conversation. Consider the example in Figure 1, from Let's Go! Public (Raux et al, 2006), a spoken dialog system that provides bus schedule information in Pittsburgh. In the first turn, the system asked for the departure location. The user responded "the airport", but this was misrecognized as "Liberty and Wood". Next, in turn 2, the system tried to explicitly confirm the departure location it heard. The user corrected the system by answering "no". The immediate reason

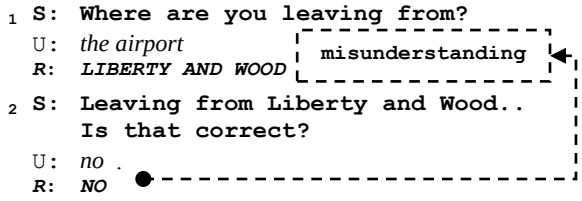

Figure 1. User responses to explicit confirmation questions can provide labels for building a confidence annotation model

for the user response in turn 2 was to allow the conversation to proceed correctly. Notice however that this interaction pattern generates additional useful information: the system now knows that it misunderstood the user in turn 1 and can use this information to refine the confidence annotator.

Spoken dialog systems should be able to successfully elicit and leverage this and other interaction patterns to continuously improve their performance, without developer supervision. For instance, we can envision a system that starts by explicitly confirming all the pieces of information it acquires from the user – many systems do this routinely. As the system collects more labels through interaction and updates its confidence annotation model, its error detection abilities improve and the system can start trusting the confidence annotation model more, and use explicit confirmations only when the confidence score is very low. Several interesting questions arise: (1) can a system make effective use of the information obtained through interaction? (2) How can a system balance its long-term knowledge elicitation goals with the short-term need to efficiently provide information to the user? (3) Could a system discover new interaction patterns that can provide labels for confidence?

We believe that implicitly-supervised learning approaches can be used in a number of other problems in spoken language interfaces (more on this in Section 7.) The work described in this paper constitutes only a starting point for a larger research program aimed at investigating the properties, advantages and limitations of this paradigm. We begin our investigation by applying the proposed approach to the confidence annotation problem. Moreover, we focus for now only on the first one of the three questions we have raised above: can a system make effective use of the information obtained through interaction to build a high quality confidence model? In future work, we plan to address the remaining questions, and to investigate the use of this paradigm in other problems.

2 Implicitly supervised learning for confidence annotation

We have already outlined the basics of using implicitly-supervised learning for building confidence annotation models. The key idea is that the system can obtain the required supervision signal by leveraging a certain pattern that occurs naturally in

conversation: in this case user responses to explicit confirmation questions. This eliminates the need for developer supervision (i.e. for manually labeling data) and in the process creates an opportunity for continuous, on-line learning. The implicitly obtained labels (**implicit labels** in the sequel) can be used in conjunction with a traditional supervised learning methodology to construct or refine a confidence annotation model.

More specifically, the implicit labels are generated automatically as follows: if the system engages in an explicit confirmation and the recognized user response was yes (or equivalent), then the previous user turn is labeled as correctly understood by the system; alternatively, if the recognized user response was no (or equivalent) the previous user turn is considered misunderstood by the system; finally if the recognized user response did not contain a positive or negative marker, no implicit label is generated. Note that the implicit labels are not noise-free. In the example from Figure 1, the user response was a simple "no", which was correctly understood by the system. In general, user responses to explicit confirmation actions extend beyond simple yes and no answers, and can also be subject to recognition errors (Krahmer et al, 2001; Bohus and Rudnicky, 2005.) As a consequence, the labels produced by this interaction pattern will not always be perfect.

The implicit labels can be characterized in terms of **accuracy** and **recall**. In this context, by **accuracy** we will refer to the accuracy of the implicit labels with respect to the reference set of manual labels. By **recall** we refer to the proportion of utterances for which this interaction pattern can generate labels (i.e. the utterances followed by an explicit confirmation and a simple user response.) Finally, there is a third factor that affects the quality of the implicitly labeled data: **the sampling bias**. Even though the proposed interaction pattern provides labels for a certain proportion of the utterances in the corpus, these implicitly labeled utterances do not constitute a random sample of the entire corpus. Rather, these are utterances that are followed by explicit confirmations, which in turn are followed by simple user responses. The underlying distribution of the features in this subset of utterances does not necessarily match the general distribution in the full set of utterances. Similarly, because this implicit labeling scheme relies on recognition of user responses, it might bias the implicit labels towards one of the two classes.

Whether or not these implicit labels are sufficient for training an accurate confidence annotation model remains an open question. In this paper, we empirically investigate this question, using corpora collected with two different spoken dialog systems.

3 Systems

The first system, Room-Line, is a telephone-based, mixed-initiative spoken dialog system that can assist users in making conference room reservations on the CMU campus (Bohus, 2007). The system has access to the live schedules of 13 conference rooms on campus, and to their characteristics, and can engage in a negotiation dialog to identify the room that best matches the user's needs.

The second system, Let's Go! Public (Raux et al, 2006), provides bus route and schedule information in the greater Pittsburgh area. Since March 2005, this system has been connected to the Pittsburgh Port Authority customer service line during non-business hours, and therefore receives a large number of calls from users with real needs.

4 Data

The RoomLine corpus consists of 484 dialogs (8037 user turns) collected in a user study in which 46 participants were asked to perform 10 scenario-based interactions with the system. The Let's Go! Public corpus consists of a subset of 617 dialog sessions (6029 utterances) collected during the first month of public operation for the system. Both corpora were orthographically transcribed, and misunderstandings were manually labeled. Table 1 shows a number of basic corpus statistics.

The RoomLine and Let's Go! Public systems used very different policies for engaging in explicit confirmations. RoomLine made this decision by comparing the confidence score of the recognized utterance against a confirmation threshold. As a result, the total number of explicit confirmations in this corpus is 1412, amounting to 17.6% of the total number of utterances (8037). In contrast, given the more adverse environment, the Let's Go! Public system used a simpler, more conservative confirmation policy: the system always explicitly confirmed every piece of information received from the user. The number of explicit confirmations in the Let's Go! Public corpus is therefore signifi-

Statistics	RoomLine	Let's Go
# of sessions	484	617
# of utterances	8037	6029
# of misunderstandings	1523	1863
% misunderstandings	18.9%	30.9%
# of explicit confirmations	1412	2594
% of explicit confirmations	17.6%	43.0%
# Implicit labels	976	1998
Implicit labels recall	10.8%	33.1%
Implicit labels accuracy	89.9%	82.5%

Table 1. Corpora statistics

cantly larger – 2594, representing 43.0% of the total number of utterances (6029).

Due to the different confirmation policies, the recall and the accuracy of the implicit labeling scheme proposed above was different in these two domains. As expected, given that explicit confirmations were more often engaged in the Let's Go! Public system, the recall of the implicit labeling scheme was significantly larger than in the RoomLine system: 33.1% versus 10.8%. At the same time, given the more adverse noise conditions and worse recognition performance in this domain, the accuracy is lower: 82.5% versus 89.9% in the RoomLine system.

5 Features

To build the confidence annotation model, we considered a large set of features extracted from different knowledge sources in the systems. Below, we give a brief overview of these features. The full feature set is presented in detail in (Bohus, 2007):

- **speech recognition features,** e.g. acoustic and language model scores; # of words and frames; word-level confidence scores generated by the recognizer; signal and noise-levels; speech-rate; etc.
- **prosody features,** e.g. various pitch characteristics such as mean, max, min, standard deviation, min and max slopes, etc.
- **lexical features,** e.g. presence or absence of the top-10 words most correlated with misunderstandings (these are system-specific.)
- **language understanding features,** e.g. number of (new / repeated) semantic slots in the parse; measures of parse-fragmentation;
- **inter-hypotheses features.** features describing differences between the top-most hypothesis from each recognizer (each system used 2 gender-specific parallel recognizers);
- **dialog management,** e.g. match-score between the recognition result and the dialog manager expectation; dialog state; etc.
- **dialog history,** e.g. # of previous consecutive non-understandings; ratio of non-understandings up to the current point in the dialog; tallied averages of the acoustic-, language-model, and parse-scores.

6 Experimental results

We used stepwise logistic regression (Myers et al. 2001) to train confidence annotation models based on the implicitly labeled portions of the RoomLine and Let's Go! Public corpora. The features described in the previous section served as independent variables in the model; the dependent (target) variable was whether or not the utterance was correctly understood by the system. The models were trained and evaluated using a 20-fold cross-validation procedure. The quality of the models was assessed in terms of mean squared error, also known as Brier score. In contrast to classification error metrics, the Brier score is a proper-scoring rule that captures both the refinement (accuracy) as well as the calibration of the confidence annotator (Cohen and Goldszmidt, 2004.)

We begin by describing results in the Let's Go! Public system, because the number of implicitly labeled training points in this corpus is larger and enables a more robust analysis.

6.1 Results in the Let's Go! Public domain

The results are illustrated in Figure 2. The Brier score for the majority baseline (i.e. always predicting the majority class) is 0.2156. The average test-set Brier score for the fully-supervised model, i.e. the model that uses the entire Let's Go! Public corpus with the manually annotated labels, is 0.1200. The proposed implicitly-supervised approach leads to an average test-set Brier score of 0.1443, closing 75% of the gap between the majority baseline and the fully-supervised model, without requiring any manually labeled data.

If a small amount of manually labeled data is available, it can be used to calibrate the implicitly-supervised model. The post-calibration step consists of training the parameters of an additional sigmoid to map the implicitly-supervised model scores into more accurate probabilities, based on the manually labeled data (Platt, 1999.) This pro-

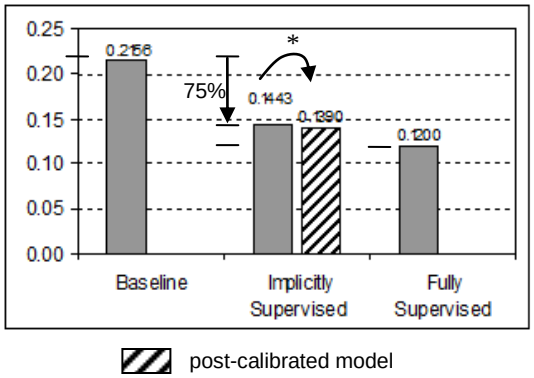

Figure 2. Implicitly- versus fully-supervised learning on Let's Go! Public data

cedure (based in our case on 100 randomly chosen labeled data-points) further increased the model's performance to 0.1390, therefore closing 80% of the gap between the baseline and fully supervised model. The difference between the un-calibrated and calibrated models is statistically significant (paired t-test, p=0.002).

The remaining performance gap between the implicitly and fully-supervised models is explained by the recall, accuracy and sampling bias of the implicit labels To better understand the effect of these factors on model performance, we constructed a number of additional models.

First, to distinguish between the effects of accuracy and recall, we constructed a model, dubbed **full-accuracy/same-recall (FA/SR)**. In training this model we only used the subset of utterances that were implicitly labeled (hence same-recall), but in conjunction with the manually obtained labels for these utterances (hence full-accuracy). The average test-set Brier score for this model was 0.1321, about half-way between the implicitly-supervised and fully-supervised models, with both differences statistically significant ($p<10^{-6}$) – see Figure 3. This result indicates that both the lack of recall and the lack of accuracy in the implicit labels contribute in roughly equal amounts to the observed performance gap.

Next, we constructed two additional models to investigate the effect of sampling bias on performance. (Recall that the subset of implicitly labeled utterances does not constitute a random sample for the entire corpus.) The first one of these models, **full-accuracy/random-same-recall (FA/RR)**, addresses the recall-bias issue and was trained with a randomly selected subset of utterances that has the same recall (size) as the implicitly labeled subset (hence random-same-recall). The second model, **random-same-accuracy/ same-recall (RA/SR)**, addresses the accuracy-bias issue. This model uses the utterances that were implicitly labeled (hence same-recall); the training labels were however constructed by starting from the reference labels and randomly altering them to attain the same accuracy level as the implicit labels have.

The performance of the full-accuracy/random-same-recall model, 0.1239, places it closer to the fully-supervised model (0.1200) than to the full-accuracy/same-recall-model (0.1321) – see Figure 3. Both differences are statistically significant in a paired t-test. The larger difference to the full-accuracy/same-recall model seems to indicate that the recall bias does affect performance in this case. On the other hand, the random-same-accuracy/ same-recall model performs similarly to the implicitly supervised model, in fact slightly worse (0.1475 versus 0.1443, no statistically significant

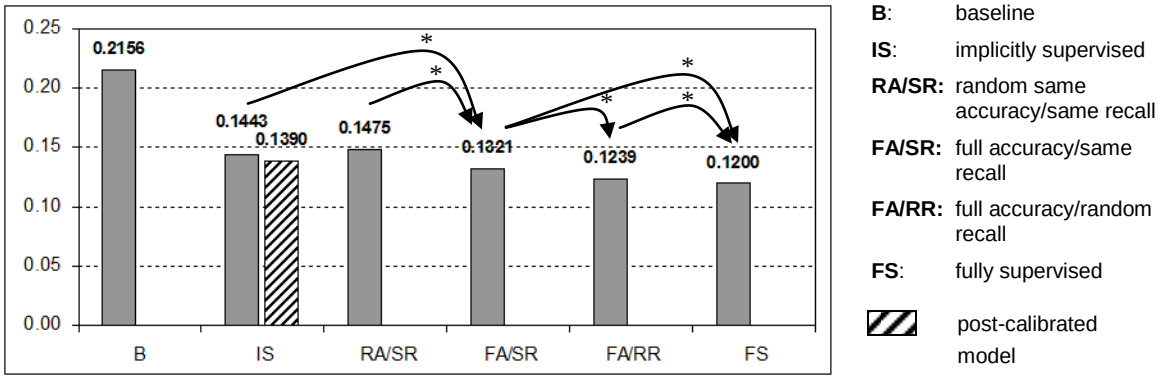

Figure 3. Implicitly- versus fully-supervised learning performance gap decomposition in Let's Go! Public domain (arrows with stars mark statistically significant differences, p<0.001)

difference detected). This result indicates that, at least in the Let's Go! Public system, the proposed implicitly generated labels do not exhibit a detrimental accuracy bias.

On a final note, recall that in Figure 2 we have seen that the implicitly-supervised approach closes 75% of the gap between the majority baseline and a fully-supervised approach (using the whole corpus). A comparison with the full-accuracy/random-same-recall model is more informative, because this model uses the same amounts of labeled data. Correcting for sample bias represents a difficult and interesting research problem (Zhang and Rudnicky, 2006). At the same time, we can easily envision using more data (since we don't need to manually label it.) As more data becomes available, the full-accuracy/random-same-recall model will eventually reach the performance of the fully supervised model. When compared to this model, the proposed implicitly-supervised approach closes 78% of the performance gap; the post-calibrated model closes 84% of this gap.

6.2 Results in the RoomLine domain

We now shift our attention to the RoomLine domain. Here, due to the more optimistic confirmation policy, the recall of the proposed implicit labeling scheme is lower: 10.8%. At the same time, due to better environmental conditions and less recognition errors, the accuracy is higher: 89.9%.

The results in this domain are illustrated in Figure 4. The implicitly-supervised approach again attains a significant improvement over the majority baseline. The relative improvement is smaller than the one attained in the Let's Go! Public domain. On the RoomLine corpus, the implicitly-supervised approach closes only 48% of the gap to the fully-supervised model; the post-calibrated model performs slightly better, but the improvement is not statistically significant. When compared to the full-accuracy/random-same-recall model, the implicitly supervised approach closes 59% of the gap (vs 78% in the Let's Go! Public domain.)

The lower performance on the RoomLine domain was expected due to the more optimistic confirmation policy and the resulting lower recall of the implicit labeling scheme. Overall, the RoomLine corpus contains 977 implicitly labeled training points, while the Let's Go! Public corpus contains more than double that amount. In the ideal case, in order to build a confidence annotation model using the proposed implicitly-supervised approach we would like the system to start with an always-confirm policy, like in the Let's Go! Public system. The full-accuracy/same-recall model (FA/SR in Figure 4), confirms that a significant part of the remaining performance gap is indeed explained by the lower recall. At the same time, part of the remaining performance gap is also explained by the lack of accuracy. This is somewhat surprising, since the accuracy is higher than in the Let's Go! Public domain. A possible explanation is that, when only small amounts of data are available for training, and/or when the class marginals are more skewed, precision plays a more important role.

Finally, the random-same-accuracy/same-recall and full-accuracy/random-same-recall models reveal that there is no detrimental sampling or recall bias in this domain. Like before, as the amount of training data increases, we can expect the gap between the full-accuracy/same-random-recall and fully-supervised model to decrease; further per-

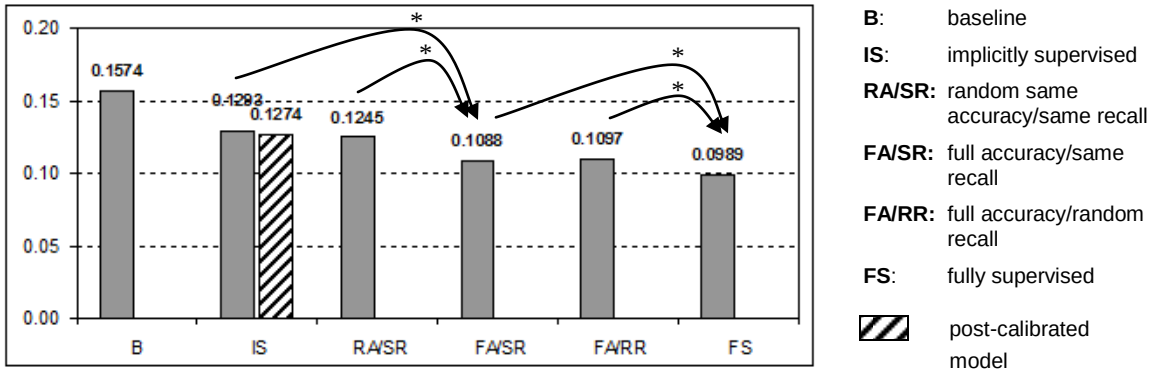

Figure 4. Implicitly- versus fully-supervised learning performance gap decomposition in RoomLine domain (arrows with stars mark statistically significant differences, $p<0.001$)

formance gains for the implicitly-supervised model are therefore expected, as we increase the dataset size. Experiments in which we trained the models using increasingly larger amounts of implicitly-labeled training data corroborate this conjecture (more details are presented in Appendix A.)

7 Discussion and future plans

While the empirical results we described in the previous section are very encouraging, they represent only a first step towards understanding the properties, advantages and limitations of the proposed implicitly-supervised paradigm.

So far, we have only performed a batch mode evaluation. However, apart from eliminating the need for a manually labeled corpus, a second important advantage of the implicitly-supervised approach is that it facilitates online learning and adaptation. The next question therefore is: how can a system engage in explicit confirmations in pursuit of its learning goals but without significantly disrupting the interaction? This is a control problem, where the system must balance the benefits of gaining knowledge via explicit confirmations against the costs potentially incurred by the user.

To some extent, dialog managers already have to solve similar trade-offs when deciding between different confirmation strategies, for instance between explicit or implicit confirmation. Explicit confirmations take an extra dialog turn, but the system has a better chance of understanding the follow-up user response, especially if the information to be confirmed is incorrect (Krahmer et al, 2001; Bohus and Rudnicky, 2005.) Typically, the costs are assumed to be known and are immediate. Solutions to these trade-off problems range from hard-coded heuristics to various offline corpus-based methods. In an online implicitly-supervised approach, the additional learning goals change the nature of the problem in two different ways. First, system actions not only create immediate dialog costs, but also produce knowledge that can be used to improve future performance. To address this new trade-off, the system must be able to assess the long-term benefits of the knowledge that stands to be gained. Secondly, in order to provide an online solution, systems should be able to continuously monitor their current performance and adjust their control policies, as their models improve.

Finally, another interesting question regards the knowledge-producing interaction pattern itself. In the experiments discussed above, the pattern consisted of user responses to system confirmation questions. Intuitively, other informative patterns could be found. For instance, if in a certain segment the dialog advances normally towards its goals, and no non-understandings occur, we might consider all those user turns correctly understood by the system. Alternatively, if a certain concept is corrected by the user at a later point in the dialog, we might mark the utterance from which the system extracted the first value for that concept as incorrect. We believe that an interesting avenue for future research is to develop techniques that allow systems to automatically discover such knowledge-producing interaction patterns.

The central idea in the proposed implicitly-supervised learning paradigm is therefore to acquire knowledge online, by discovering, eliciting and leveraging natural patterns that occur in interaction as a by-product of the collaboration between the system and an invested user. This paradigm can eliminate the need for developer supervision and can enable fast online adaptation and learning. We conjecture that it can supplement and or even provide a strong alternative to existing learning approaches, and enable significant autonomous learning in interactive systems.

The use of implicit feedback and human supervision for labeling, learning or adaptation purposes appears before in a number of other areas, like information retrieval (Brown and Claypool, 2003; Shen et al, 2005), image labeling (von Ahn and Dabbish), meeting segmentation (Banerjee and Rudnicky, 2007). To our knowledge, the work described in this paper is the first effort in learning from implicit supervision in the context of conversational spoken language interfaces. While in this paper we have focused only on one learning problem (i.e. building confidence annotation models), we believe that the proposed implicitly-supervised paradigm can be applied to a number of other problems in conversational spoken language interfaces. In fact, we have already developed and will soon report on an implicitly-supervised approach for learning how to automatically correct non-understanding errors in a spoken dialog system.

8 Conclusion

In this paper, we have proposed the use of an implicitly-supervised approach for learning in spoken language interfaces and have applied it for constructing confidence annotation models. Previous supervised learning solutions (Litman et al, 1999; Carpenter et al, 2001; San-Segundo et al, 2001; Hazen et al, 2002; Hirchberg et al, 2004.) rely on pre-existing, in-domain, manually labeled data and lead to static solutions. In contrast, the proposed approach does not require developer supervision. Instead, the system obtains the supervision signal from follow-up user responses to the system's explicit confirmation questions. In effect, the system learns from its own experiences.

We evaluated the proposed approach in two different dialog domains: RoomLine and Let's Go! Public. Empirical results confirm that a system can indeed successfully leverage interaction patterns to automatically construct a confidence annotation model that performs similarly to a fully-supervised model. The experiments we have reported here represent only a first step towards a fuller understanding of the proposed implicit-learning paradigm. The encouraging results we have obtained on the confidence annotation task point towards what we believe to be a very interesting research avenue. We conjecture that the proposed approach can be applied to address a number of other problems in conversational spoken language interfaces, and in interactive systems in general. Ultimately, we hope that it will enable the development of autonomously self-improving systems.

Acknowledgements

The authors would like to thank Antoine Raux for helpful discussions and suggestions during the early stages of this work, and Maxine Eskenazi and Alan W Black for making the Let's Go! Public corpus available for research purposes.

References

Banerjee, S., and Rudnicky, A. 2007. *Segmenting Meetings into Agenda Items by Extracting Implicit Supervision from Human Note-Taking*, in Proc. of IUI'07, Honolulu, Hawaii, USA.

Bohus, D., and Rudnicky, A. 2005. *Constructing Accurate Beliefs in Spoken Dialog Systems*, in Proc. of ASRU-2005, San Juan, Puerto Rico.

Bohus, D. 2007. *Error Awareness and Recovery in Conversational Spoken Language Interfaces*, Ph.D Thesis, Carnegie Mellon University, CS-07-124

Brown, D., and Claypool, M. 2003. *Curious Browsers: Automated gathering of implicit interest indicators by an instrumented browser*, in Workshop on Implicit Measures of User Interests and Preferences, SIGIR'2003, Toronto, Canada

Carpenter, P., Chun, J., Wilson, D., Zhang, R., Bohus, D., and Rudnicky, A. 2001. *Is this conversation on track?*, in Proc. of Eurospeech'99, Aalborg, Denmark

Cohen, I. and Goldszmidt, M., 2004 - *Properties and benefits of calibrated classifiers*. in Proc. of EMCL/PKDD. Pisa, Italy.

Hazen, T., Seneff, S., and Polifroni, J. 2002. *Recognition confidence scoring and its use in speech understanding systems*, Computer Speech and Language.

Hirchberg, J., Litman, D., and Swerts, M. 2004. *Prosodic and other cues to speech recognition failures*, Speech Communication, 2004.

Krahmer, E., Swerts, M., Theune, M., and Weegels, M., 2001. *Error Detection in Spoken Human-Machine Interaction*. International Journal of Speech Technology. 4(1): p. 19-30.

Litman, D., Walker, M., and Kearns, M. 1999. *Automatic Detection of Poor Speech Recognition at the Dialogue Level*, in Proc. of ACL'99, College Park, MD.

Myers, R. H., Montgomery, D. C., and Vining, G. 2001. *Generalized Linear Models: With Applications in Engineering and the Sciences*. Wiley series in probability and statistics, ed. Wiley-Interscience.

Platt, J. 1999. *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*, Advances in Large Margin Classifiers

Raux, A., Bohus, D., Langner, B., Black, A.W, and Eskenazi, M. 2006. *Doing Research in a Deployed Spoken Dialog Systems: One Year of Let's Go! Public Experience*, in Proc. of Interspeech'06, Pittsburgh, PA.

San-Segundo, R., Pellom, B., Hacioglu, K., and Ward, W. 2001. *Confidence Measures for Spoken Dialogue Systems*, in Proc. of ICASSP'01, Salt Lake City, UT

Shen, X., Tan, B., and Zhai, ChengXiang, 2005. *Context-Sensitive Information Retrieval using Implicit Feedback*, in Proc. of SIGIR'05, Salvador, Brazil

von Ahn, L., and Dabbish, L., 2004. *Labeling images with a computer game*, in CHI'04, New York, NY

Zhang, R. and Rudnicky, A. 2006. *A New Data Selection Approach for Semi-Supervised Acoustic Modeling*, in Proc. of ICASSP'06. Toulouse, France.

Appendix A. Performance as a function of training set size

We investigated the relationship between the performance of the implicitly-supervised confidence annotation models and the overall training set size. The results are shown in Figure 5.A for the RoomLine domain, and Figure 5.B for the Let's Go! Public domain.

In the RoomLine domain, the performance of the implicitly-supervised model does not yet reach an asymptote by the time we have considered the full training set (7537 utterances.) This result corroborates our previous conjecture that, if more data were available, further performance gains would be possible. As more data becomes available, the full-precision/random-same-recall model is guaranteed to reach the same asymptote as the fully supervised model. At the same time, we expect that the gap between the implicitly supervised method and the full-precision/random-same-recall model will stay roughly constant. As a consequence, we expect corresponding gains in the implicitly-supervised model performance.

Another interesting observation is that the random-same-precision/same-recall model closely tracks the implicitly supervised model, and the full-precision/random-same-recall model closely tracks the full-precision/same-recall model. These trends confirm that there is no detrimental sample bias (neither in terms of accuracy nor recall) in the proposed implicit learning scheme in the RoomLine data.

In the Let's Go! Public domain, the implicitly-supervised model seems to have reached a performance asymptote; this is not surprising, given the larger recall of the implicit labeling scheme in this domain. As the amount of data increases, the full-precision/random-same-recall model shows increasingly larger improvements over the full-precision/same-recall model.

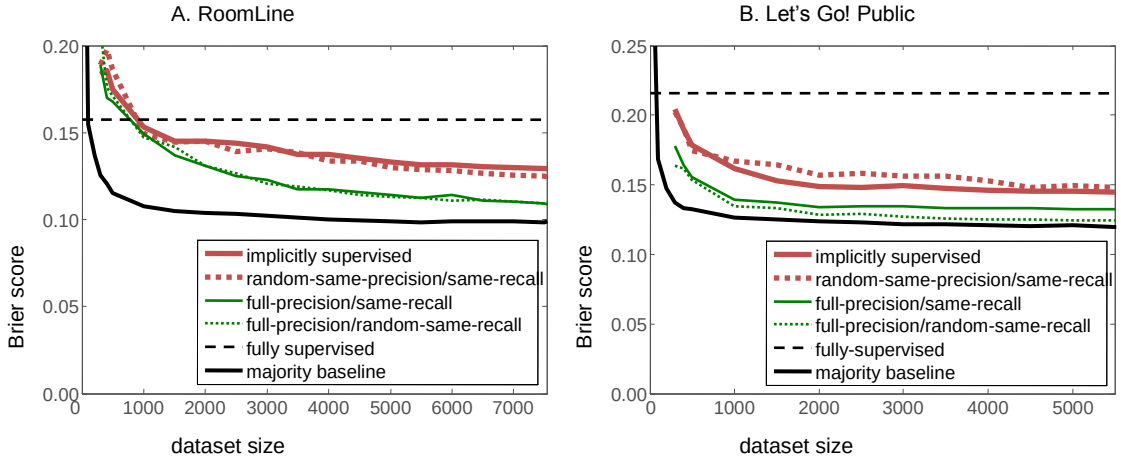

Figure 5. Implicitly supervised confidence annotation model performance as a function of training set size (in the RoomLine and Let's Go! Public domains)