Improving Classification-Based Natural Language Understanding with Non-Expert Annotation

Fabrizio Morbini and **Eric Forbell** and **Kenji Sagae**
Institute for Creative Technologies
University of Southern California
Los Angeles, CA 90094, USA
{morbini,forbell,sagae}@ict.usc.edu

Abstract

Although data-driven techniques are commonly used for Natural Language Understanding in dialogue systems, their efficacy is often hampered by the lack of appropriate annotated training data in sufficient amounts. We present an approach for rapid and cost-effective annotation of training data for classification-based language understanding in conversational dialogue systems. Experiments using a web-accessible conversational character that interacts with a varied user population show that a dramatic improvement in natural language understanding and a substantial reduction in expert annotation effort can be achieved by leveraging non-expert annotation.

1 Introduction

Robust Natural Language Understanding (NLU) remains a challenge in conversational dialogue systems that allow arbitrary natural language input from users. Although data-driven approaches are now commonly used to address the NLU problem as one of classification, e.g. (Heintze et al., 2010; Leuski and Traum, 2010; Moreira et al., 2011), where input utterances are mapped automatically into system-specific categories, the dependence of such approaches on training data annotated with semantic classes or dialogue acts creates a chicken and egg problem: user utterances are needed to create the annotated training data necessary for NLU by classification, but these cannot be collected without a working system that users can interact with.

Common solutions to this problem include the use of Wizard-of-Oz data collection, where a human expert manually provides the functionality of data-driven modules while data is collected from users, or the use of scenario authors who attempt to anticipate user input to create an initial set of training data. While these options offer practical ways around the training data acquisition problem, they typically require substantial work from system experts and provide suboptimal solutions: data-driven approaches work best when utterances in the training data are drawn from the same distribution as those encountered in actual system use, but the conditions under which training data is collected (a human expert filling in for systems modules, or a human expert generating possible user utterances) are quite different from those where users interact with the final system. High quality results are often obtained through an iterative process where an initial training set is authored by a scenario designer, but NLU resources are gradually updated based on real user data over time (Gandhe et al., 2011). Although this can ultimately produce training data composed primarily of real user utterances, and therefore result in better performance from data-driven models, an expert annotator is required to perform manual classification of user utterances. This is a laborious process that assumes availability and willingness of the annotator for as long as it takes to collect enough user utterances, which may range from weeks to months or even years, depending on the size of the domain and the number and type of utterance categories.

The main question we address is whether annotation by non-experts can be leveraged to speed up utterance classification and lower its cost. We present a technique that frames the annotation of training data as a human intelligence task suitable for crowdsourcing. Although there are similarities between our technique and active learning (e.g. see (Gambck et al., 2011)), an important difference is that our technique does not reduce the annotation effort by reducing the size of the data to be labeled, but by casting the annotation task into a simpler problem. This allows us to take advantage of the entire data generated by the users. Through an experiment with a conversational dia-

Proceedings of the SIGDIAL 2014 Conference, pages 69–73,
Philadelphia, U.S.A., 18-20 June 2014.

logue system deployed on the web, we show that a dramatic improvement in the quality of NLU can be achieved with non-expert data annotation, reducing the time required of an expert annotator by 70%.

2 Improving understanding with data

Our approach for creating accurate utterance classifiers for NLU in conversational dialogue systems is based on a simple strategy, which we describe next in general terms. NLU is assumed to be performed through multiclass classification.

The first step is to create a small initial training dataset T_0 either through Wizard-of-Oz data collection or by generation of utterances by a system developer or content author. This training set is used to train a NLU model M_0. Although this model is likely to be inadequate, it allows users to interact with an initial version of the system. As input utterances are collected from real users, these utterances are annotated with their desired NLU output labels. Periodically, at time i, we add to the initial training dataset T_0 the annotated user utterances accumulated up to that point. We train a new NLU model M_i using this augmented training set, T_i.[1] We also keep aside a small fraction of utterances to test the performance of the NLU models, that is, at each time i we also have an evaluation set E_i and the union of E_i and T_i is the entire set of user utterances collected up to time i. As more utterances are added and annotated, an NLU model M_i is expected to surpass the initial model M_0. In general, we replace the running NLU model M_r whenever we have a better performing M_i model. This straightforward process can be used to obtain increasingly more accurate language understanding, at the cost of data annotation in the form of labelling utterances with categories that are defined according to the needs of the specific system and the specific domain. The categories may be based on dialogue acts, e.g. (Core and Allen, 1997; Bunt et al., 2010), user information needs, e.g. (Moreira et al., 2011), or stand in for entire semantic frames, e.g. (DeVault and Traum, 2013). The technical nature of the task of categorizing utterances in schemes such as these usually means that substantial time is required of an expert annotator.

2.1 Annotation as a human intelligence task

Although the task of annotating NLU training data involves assigning categories with technical definitions to utterances, and therefore would appear to require knowledge of these technical definitions, in fact the task requires primarily the type of language understanding that is common to all native speakers of a given language. Our main hypothesis is that this annotation can be structured as a trivial task that requires no specific expertise, and that annotations performed this way can have a substantial impact on the quality of utterance classification. We define the NLU annotation task as follows.

Before annotation begins, each utterance category in the system is associated with one or more canonical utterance(s) that capture the meaning and communicative intent of that category. These canonical utterances allow for trivial generation of surface text from NLU categories by simple lookup.

Given a set of dialogues between the system and individual users, each dialogue is split into segments. Segments are typically composed of a user utterance and the preceding system utterance. In cases where users produce multiple utterances without intervening system utterances, segments may consist of a system utterance followed by consecutive user utterances, so that there is exactly one segment corresponding to each user utterance u in the dialogue and each segment starts with a system utterance and ends with u.

Annotators work on one dialogue at a time, and within a dialogue they are presented with one segment at a time, following the order in which the utterances appear in the dialogue. Assignment of a system category to the utterance targeted in a segment is a multiple choice task, involving n choices, one of which is *none of the above*. The remaining $n-1$ choices are generated automatically by running the current NLU utterance classifier on the target utterance, and generating the canonical surface realizations of each of the $n-1$ highest scoring categories. The annotator is then asked to choose the option containing language that means the same as the targeted utterance in the segment. The choice of n depends on the quality of the current NLU model, and must be made to balance the length of the list of options with the expected number of times the appropriate choice is not in the list of options. The more *none of the above* annotations are performed, the less effective the annotation exercise will be in reducing the effort required by an annotation expert.

The utterances annotated as *none of the above* are given to an expert for annotation. The impact

[1] For every time i and j with $i < j$ it holds that $T_i \subseteq T_j$.

System Hello and welcome.
...
User Are you married?
System Yeah, forty-four years this September. I can't believe Linda's stuck with me this long.
System Okay, let's get to it. Why are you here today?
User I've been having trouble sleeping lately.
System This must be hard for you.
System Are you feeling stressed at all?
User Yes, quite a bit
System So, I've got some questions I need to ask you, to get to know you better.
System Just double checking, you haven't been sleeping good. Is that right?
User That's right
System Have you been having any nightmares?
User Every once in a while, but mostly I can't stay asleep
...

Figure 1: Excerpt of a sample interaction with a Simcoach character over the web. User input is typed into a text box on a web page.

of this procedure on how quickly user utterances are annotated as training examples for the NLU and how much expert time is saved depends on the quality of the initial classification model and the choice of n. These are practical factors that are expected to vary greatly from one dialogue system to another. In the next section, we present one application of this procedure to an existing conversational dialogue system deployed on the web, and show examples of dialogue segments and annotation options.

3 Experiment

To test our hypothesis that language understanding can be improved with much reduced expert effort, we applied the framework described above to a system that implements a conversational character that talks with users about issues relating to mental and behavioral disorders and presents health care options. The system is publicly accessible at http://www.simcoach.org, and receives traffic on the order of one hundred users per week. Of these, about one quarter engage the system in a meaningful dialogue with multiple turns, with the dialogues containing on average 16 user utterances. Because our process depends crucially on user traffic to generate data for annotation, a web-accessible system is ideally suited for it. An excerpt from a typical interaction with the system is shown in Figure 1. The system and the NLU classifier based on Maximum Entropy models (Berger et al., 1996) are described respectively in (Rizzo et al., 2011) and (Sagae et al., 2009).

3.1 Data collection

Starting with an initial system deployed with an NLU model trained with data generated by an author attempting to anticipate user behavior, we applied the approach described in section 2 to improve NLU accuracy over a period of approximately five months. The initial accuracy of the NLU classifier was 62%, measured as the number of utterances classified correctly divided by the total number of user utterances. This accuracy figure was obtained only after the five months of data annotation, using the heldout set of manually annotated dialogues.

Although the data annotation procedure as described in section 2 could in principle be performed continuously as user data come in, we instead performed all of our annotation in three rounds, the first consisting of approximately 2,000 user utterances, the second one month later, consisting of an additional 1,000 utterances. The last round, collected about two months later, contained about 2,000 utterances. We used five annotators[2] working in parallel, and the average speed of each annotator exceeded 500 utterances per hour.

The total number of NLU utterance classes in the system is 378, although only 120 classes were used by annotators in all rounds of annotation to cover all of the utterances collected[3]. In our annotation exercise we set the number of multiple choice items at $n = 6$, including 5 choices generated from categories chosen by the NLU classifier, and one *none of the above* choice. Figure 2 shows a sample dialogue segment with the corresponding multiple choice items. During annotation, clicking on a multiple choice item advances the annotation by presenting the next segment containing a user utterance to be annotated.

3.2 Results

Of the utterances in the three rounds of data collection, respectively 29%, 34% and 17% were marked by annotators as *none of the above*. These were given to a developer of the NLU system who assigned a category to each of them. In this expert annotation step the choice is not restricted to a small set of options, and may be any of the categories in the system. Given this rate of use of

[2]The non-expert annotators belonged to the same team that developed the system but did not participate in the development of the NLU module and the NLU classes used in the particular dialogue system used.

[3]This difference is a further evidence of the difficulty of correctly anticipating how the end users will interact with the dialogue system.

System Okay, let's get to it. Why are you here today?

User I've been having trouble sleeping lately.

Which of the following options correspond most closely to the last user utterance? If none of them have the same general meaning as the user utterance, select "none of the above."

(a) I have been in a bad mood lately
(b) I have nightmares often
(c) I haven't been sleeping well
(d) My family is worried about me
(e) I eat too much
(f) None of the above

Figure 2: Example of a dialogue segment with corresponding multiple choice items. The annotation task consists of choosing the item that has approximately the same meaning and communicative intent as the targeted utterance (the user utterance).

the *none of the above* category, the need for expert annotation is not eliminated, but the amount of expert effort necessary is reduced by over 70%.

The NLU classification accuracy figures obtained after each round of annotation are shown in Table 1. In the table, *Our Approach* represents the results obtained by the technique described here. A large improvement is observed after the first round of annotation, with a more modest improvement observed after the other two rounds. The initial jump in accuracy after round 1 is explained by the fact that the initial model based on a system author's expectation of what users may say to the system (approximately 3,000 utterances) is improved using utterances that users did in fact produce in real interactions with the system. Clearly, a more well-matched distribution of utterances in the training data produces higher accuracy.

To assess the value of our approach, we compare it with two other reasonable experimental conditions: a baseline where only expert annotation is used (*Expert Only*), and a condition where no expert annotation is used (*No Expert*). The *Expert Only* condition is meant to represent what can be achieved with the same workload for the expert used in *Our Approach*. This is achieved by random selection of user utterances to create a set with the same number of utterances set aside for expert annotation in *Our Approach*. The expert then annotates each of these utterances to create training data. For the *No Expert* condition, we used only utterances annotated by non-experts, leaving out completely utterances labeled as *none of the above*. Both *Expert Only* and *No Expert* conditions achieve significantly lower performance than the approach described here. This indicates that expert annotation is important, but also that cheap and fast non-expert annotation can provide substantial improvements to NLU.

	NLU accuracy after each annotation round [%]			
	Base	1st round	2nd round	3rd round
Our Approach	62	**70**	**73**	**78**
Expert Only	62	64	68	70
No Expert	62	64	65	71

Table 1: NLU accuracy obtained using the initial training dataset T_0, after one round of annotation with T_1 (2,013 utterances), after two rounds of annotation with T_2 (additional 948 utterances), and after three rounds with T_3 (additional 1806 utterances). Accuracy is estimated on the same heldout set of dialogues E_3 for all conditions, accounting for roughly 10% of the annotated data.

4 Conclusion

We described a framework for annotation of training data by non-experts that can provide dramatic improvements to natural language understanding in dialogue systems that perform NLU through utterance classification. Our approach transforms the annotation NLU training data into a task that can be performed by anyone with language proficiency. Annotation is structured as a simple multiple choice task, easily delivered over the web.

Using our approach with a conversational character on the web, we improved NLU accuracy from 62% to 78% using only less than 30% of the effort it would be required of an expert to annotate data without non-expert annotation.

Acknowledgments

We thank Kelly Christoffersen, Nicolai Kalisch and Tomer Mor-Barak for data annotation and updates to the SimCoach system, David Traum for insightful discussions, and the anonymous reviewers. The effort described here has been sponsored by the U.S. Army. Any opinions, content or information presented does not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, March.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an iso standard for dialogue act annotation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Mark G. Core and James F. Allen. 1997. Coding dialogues with the DAMSL annotation scheme. In David Traum, editor, *Working Notes: AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, Menlo Park, California. AAAI, American Association for Artificial Intelligence.

David DeVault and David Traum. 2013. A method for the approximation of incremental understanding of explicit utterance meaning using predictive models in nite domains. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA, June.

Bjrn Gambck, Fredrik Olsson, and Oscar Tckstrm. 2011. Active learning for dialogue act classification. In *INTERSPEECH*, pages 1329–1332. ISCA.

Sudeep Gandhe, Michael Rushforth, Priti Aggarwal, and David Traum. 2011. Evaluation of an integrated authoring tool for building advanced question-answering characters. In *12th Annual Conference of the International Speech Communication Association (InterSpeech 2011)*, Florence, Italy, August.

Silvan Heintze, Timo Baumann, and David Schlangen. 2010. Comparing local and sequential models for statistical incremental natural language understanding. In Raquel Fernández, Yasuhiro Katagiri, Kazunori Komatani, Oliver Lemon, and Mikio Nakano, editors, *SIGDIAL Conference*, pages 9–16. The Association for Computer Linguistics.

Anton Leuski and David R. Traum. 2010. Practical language processing for virtual humans. In *Twenty-Second Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-10)*.

Catarina Moreira, Ana Cristina Mendes, Luísa Coheur, and Bruno Martins. 2011. Towards the rapid development of a natural language understanding module. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents*, IVA'11, pages 309–315, Berlin, Heidelberg. Springer-Verlag.

Albert A. Rizzo, Belinda Lange, John G. Buckwalter, E. Forbell, Julia Kim, Kenji Sagae, Josh Williams, Barbara O. Rothbaum, JoAnn Difede, Greg Reger, Thomas Parsons, and Patrick Kenny. 2011. An intelligent virtual human system for providing healthcare information and support. In *Studies in Health Technology and Informatics*.

Kenji Sagae, Gwen Christian, David DeVault, and David R. Traum. 2009. Towards natural language understanding of partial speech recognition results in dialogue systems. In *Short Paper Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT) 2009 conference*.