A method for the approximation of incremental understanding of explicit utterance meaning using predictive models in finite domains

David DeVault and **David Traum**
Institute for Creative Technologies, University of Southern California,
12015 Waterfront Dr., Playa Vista, CA 90094 USA
{devault,traum}@ict.usc.edu

Abstract

This paper explores the relationship between explicit and predictive models of incremental speech understanding in a dialogue system that supports a finite set of user utterance meanings. We present a method that enables the approximation of explicit understanding using information implicit in a predictive understanding model for the same domain. We show promising performance for this method in a corpus evaluation, and discuss its practical application and annotation costs in relation to some alternative approaches.

1 Introduction

In recent years, there has been a growing interest among researchers in methods for incremental natural language understanding (NLU) for spoken dialogue systems; see e.g. (Skantze and Schlangen, 2009; Sagae et al., 2009; Schlangen et al., 2009; Heintze et al., 2010; DeVault et al., 2011a; Selfridge et al., 2012). This work has generally been motivated by a desire to make dialogue systems more efficient and more natural, by enabling them to provide lower latency responses (Skantze and Schlangen, 2009), human-like feedback such as backchannels that indicate how well the system is understanding user speech (DeVault et al., 2011b; Traum et al., 2012), and more interactive response capabilities such as collaborative completions of user utterances (DeVault et al., 2011a), more adaptive handling of interruptions (Buschmeier et al., 2012), and others.

This paper builds on techniques developed in previous work that has adopted a *predictive* approach to incremental NLU (DeVault et al., 2011a). On this approach, at specific moments while a user's speech is in progress, an attempt is made to predict what the full meaning of the complete user utterance will be. Predictive models can be contrasted with *explicit* approaches to incremental NLU. We use the term *explicit* understanding to refer to approaches that attempt to determine the meaning that has been expressed explicitly in the user's partial utterance so far (without predicting further aspects of meaning to come). Explicit understanding of partial utterances can be implemented using statistical classification or sequential tagging models (Heintze et al., 2010).

Both predictive and explicit incremental NLU capabilities can be valuable in a dialogue system. Prediction can support specific response capabilities, such as system completion of user utterances (DeVault et al., 2011a) and reduced response latency.[1] However, explicit models support additional and complementary capabilities. For instance, depending on the application domain (Heintze et al., 2010) and on the individual utterance (DeVault et al., 2011b), it may be difficult for a system to predict a user's impending meaning with confidence. Nevertheless, it may often be possible for systems to determine the meaning of what a user has said so far, and to take action based on this partial understanding. As one example, items in a user interface could be highlighted when mentioned by a user (Buß and Schlangen, 2011). Another capability would be to provide grounding feedback, such as verbal back-channels or head nods (in embodied systems), to indicate when the system is understanding the user's meaning (Traum et al., 2012). Explicit utterance meanings also allow a system to distinguish between meaning that has been expressed and meaning that is merely implied or inferred, which may be less reliable. In the near future, as incremental processing capabilities in dialogue systems grow, it may prove valuable for dialogue systems to combine both predictive and explicit incremental understanding capabilities.

In this paper, we present a technique for approximating a user's explicit meaning using an existing predictive understanding framework (DeVault et al., 2011a). The specific new contributions in this paper are (1) to show that

[1] A simple approach to reducing response latency is to begin to plan a response to the predicted meaning while the user is still speaking.

Proceedings of NAACL-HLT 2013, pages 1092–1099,
Atlanta, Georgia, 9–14 June 2013.

an estimate of a user's explicit utterance meaning can be derived from this kind of predictive understanding model (Section 2); (2) to quantify the performance of this new method in a corpus evaluation (Section 3); (3) to provide concrete examples and discussion of the annotation costs associated with implementing this technique, in relation to some alternative approaches to explicit understanding (Section 4). Our results and discussion show that the proposed method offers promising performance, has relatively low annotation costs, and enables explicit and predictive understanding to be easily combined within a dialogue system. It may therefore be a useful incremental understanding technique for some dialogue systems.

2 Technical Approach and Data Set

In Sections 2.1-2.3, we briefly summarize the data set and approach to predictive incremental NLU (DeVault et al., 2011a) that serves as the starting point for the new work in this paper. Sections 2.4 and 2.5 present our new approach to explicit understanding based on this approach.

2.1 Data set

For the experiments reported here, we use a corpus of user utterances collected with the SASO-EN spoken dialogue system (Hartholt et al., 2008; Traum et al., 2008). Briefly, this system is designed to allow a trainee to practice multi-party negotiation skills by engaging in face to face negotiation with virtual humans. The scenario involves a negotiation about the possible re-location of a medical clinic in an Iraqi village. A human trainee plays the role of a US Army captain, and there are two virtual humans that he negotiates with: Doctor Perez, the head of an NGO clinic, and a local village elder, al-Hassan. The captain's main objective is to convince the doctor and the elder to move the clinic out of an unsafe marketplace area.

The corpus used for the experiments in this paper includes 3,826 training and 449 testing utterances drawn from user dialogues in this domain. The corpus and its semantic annotation are described in (DeVault et al., 2010; DeVault et al., 2011a). All user utterances have been audio recorded, transcribed, and manually annotated with the correct NLU output frame for the entire utterance. (We discuss the cost of this annotation in Section 4.) Each NLU output frame contains a set of attributes and values that represent semantic information linked to a domain-specific ontology and task model (Traum, 2003). Examples of the NLU output frames are included in Figures 2, 3, and 5.

2.2 Predictive incremental NLU

This approach uses a predictive incremental NLU module, mxNLU (Sagae et al., 2009; DeVault et al., 2011a), which is based on maximum entropy classification. The approach treats entire individual frames as output classes, and extracts input features from partial ASR results. To define the incremental understanding problem, the audio of the utterances in the training data were fed through an ASR module, PocketSphinx (Huggins-Daines et al., 2006), in 200 millisecond chunks, and each partial ASR result produced by the ASR was recorded. Each partial ASR result then serves as an incremental input to mxNLU. NLU is predictive in the sense that, for each partial ASR result, the task of mxNLU is to produce as output the *complete* frame that has been associated by a human annotator with the user's *complete* utterance, even if that utterance has not yet been fully processed by the ASR.

The human annotation defines a finite set $\mathcal{S} = \{S_1, ..., S_N\}$ of possible NLU output frames, where each frame $S_i = \{e_1, ..., e_n\}$ is a set of key-value pairs or *frame elements*. For notation, a user utterance u generally creates a sequence of m partial ASR results $\langle r_1, ..., r_m \rangle$, where each ASR result r_j is a partial text such as *we need to move*. Let G_u denote the correct (or "gold") frame for the complete utterance u. For each result r_j and for each complete frame S_i, the maximum entropy model provides $P(G_u = S_i | r_j)$. The NLU output frame S_j^{NLU} is the complete frame for which this probability is highest.

2.3 Performance of predictive incremental NLU

The performance of this predictive incremental NLU framework has been evaluated using the training and test portions of the SASO-EN data set described in Section 2.1. Performance is quantified by looking at precision, recall, and F-score of the frame elements that compose the predicted (S_j^{NLU}) and correct (G_u) frames for each partial ASR result. When evaluated over all the 5,736 partial ASR results for the 449 test utterances, the precision/recall/F-Score of this predictive NLU, in relation to the complete frames, are 0.67/0.47/0.56, respectively. When evaluated on only the ASR results for complete test utterances, these scores increase to 0.81/0.71/0.76, respectively.

2.4 Assigning probability to frame elements

An interesting question is whether we can use this model to attach useful probabilities not only to complete predicted frames but also to the individual frame elements that make up those frames. To explore this, for each partial ASR result r_j in each utterance u, and for each frame element e in SASO-EN, let us model the probability that e will be part of the correct frame for the complete utterance as:

$$P(e \in G_u | r_j) = \sum_{S_i : e \in S_i} P(G_u = S_i | r_j) \qquad (1)$$

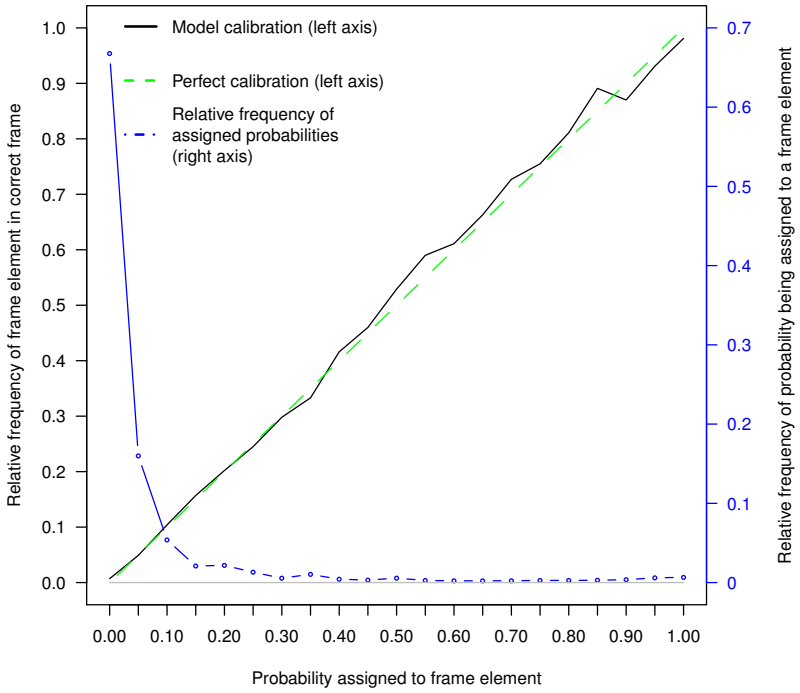

Figure 1: Calibration of frame element probabilities.

This method derives the probability of frame elements from the probabilities assigned to the possible frames that contain them. Computing this sum is straightforward in a finite semantic domain such as SASO-EN.

We computed this probability for all frame elements e and all partial ASR results r_j in our test set, yielding approximately 478,000 probability values. We grouped these probability values into bins of size 0.05, and calculated the frequency with which the frame elements in each bin were indeed present in the correct frame G_u for the relevant utterance u. The results are presented in Figure 1, which shows that the probability values derived from Equation (1) are relatively "well calibrated", in the sense that the relative frequency with which a frame element is in the final frame is very close to the numeric probability assigned by Equation (1). The figure also shows how frequently the model assigns various probability ranges to frame elements (blue dotted line, plotted against the secondary right axis). Note that most frame elements are assigned very little probability for most partial ASR results.

We conclude from these observations that the probabilities assigned by (1) could indeed carry useful information about the likelihood that individual key values will be present in the complete utterance meaning.

2.5 Selecting probable frame elements

In exploring the model of frame element probabilities given in Equation (1), we observed that often the reason a frame element has lower probability, at a given point within a user utterance, is that it is a *prediction* rather than something that has been expressed explicitly. Building on this observation, our technique for estimating the user's explicit meaning uses a probability threshold to select those individual frame elements which are most likely to be in the frame for a complete utterance, according to the predictive model. That is, at each partial result r_j, we estimate the user's explicit meaning using a constructed frame:

$$S_j^{\text{SUB}} = \{e | P(e \in G_u | r_j) \geq \tau\} \quad (2)$$

This approximation could work well if, in practice, the most probable frame elements prove to match fairly closely the user's non-incremental utterance meaning at the point this frame is constructed. We evaluate this in the next section.

Note that, in general, the returned subset of frame elements may not be identical to any complete frame $S_i \in \mathcal{S}$; rather it will correspond to parts of these complete frames or "subframes".

3 Performance Evaluation

To evaluate this technique, we constructed subsets of frame elements or "explicit subframes" using Equation (2) and various minimum probability thresholds τ for partial ASR results in our test set. We then compared the resulting subframes both to the final complete frame G_u for each utterance u, and also to manually annotated sub-

Explicit subframe (with frame element probabilities)	**Predicted complete frame**	**Annotated subframe**
Partial ASR result: *hello*		
0.813 <S>.sem.speechact.type greeting	<S>.sem.speechact.type greeting <S>.addressee doctor-perez	<S>.sem.speechact.type greeting
Partial ASR result: *hello elder*		
0.945 <S>.sem.speechact.type greeting 0.934 <S>.addressee elder-al-hassan	<S>.sem.speechact.type greeting <S>.addressee elder-al-hassan	<S>.sem.speechact.type greeting <S>.addressee elder-al-hassan

Figure 2: Explicit subframes and predicted complete frames for two partial ASR results in a user utterance of *hello elder*.

frames that represent human judgments of explicit incremental utterance meaning.

To collect these judgments, we hand-annotated a word-meaning alignment for 50 random utterances in our test set.[2] To perform this annotation, successively larger prefixes of each utterance transcript were mapped to successively larger subframes of the full frame for the complete utterance. The annotated subframes for each utterance prefix were selected to be *explicit*; they include only those frame elements that are explicitly expressed in the corresponding prefix of the user's utterance. (We discuss the cost of this annotation in Section 4.)

We provide a simple concrete example in Figure 2. This example shows two partial ASR results during an utterance of *hello elder* by a user. For each partial ASR result, three frames are indicated horizontally. At the right, labeled "Annotated subframe", we show the human judgment of explicit incremental utterance meaning for this partial utterance. Our human judge has indicated that the word *hello* corresponds to the frame element `<S>.sem.speechact.type greeting`, and that the words *hello elder* correspond to an expanded frame that includes the frame element `<S>.addressee elder-al-hassan`.

At the left, labeled "Explicit subframe", we show the subframe selected by Equation (2) for each partial ASR result, with threshold $\tau = 0.5$. A relevant background fact for this example is that in this scenario, the user can generally address either of two virtual humans who are present, Doctor Perez or Elder Al-Hassan. After the user has said *hello*, the frame element `<S>.sem.speechact.type greeting` is assigned probability 0.813 by Equation (1), and only this frame element appears in the explicit subframe.

In the middle, labeled "Predicted complete frame", the figure also shows the full predicted frame from mxNLU at each point. After the user has said *hello*, the full predicted output includes an additional frame element, `<S>.addressee doctor-perez`, indicating a prediction that the addressee of this user utterance will be Doctor Perez rather than Elder al-Hassan. However, the probability assigned to this prediction by Equation (1) is less than 0.5, and so this predicted frame element is excluded from the explicit subframe. And indeed, this is the correct *explicit* representation of the meaning of *hello* in this system.

This simple example illustrates how our proposed technique can enable a dialogue system to have access to both explicit and predicted utterance meaning as a user's utterance progresses. An excerpt from a more complex utterance is given in Figure 3. This example shows incremental outputs for two partial ASR results during a user utterance of *we will provide transportation at no cost*. In this example, the explicit subframe for *we will* includes frame elements that convey that the captain (i.e. the user) is promising to do something. This subframe does not exactly match the human judgment of explicit meaning at the right, which does not include at this point the `<S>.sem.agent captain-kirk` and `<S>.sem.type event` frame elements. However, the explicit subframe more closely matches the human judgment than does the predicted complete frame from mxNLU (middle column), which includes an incorrect prediction that the captain is promising to deliver medical supplies (represented by the key values `<S>.sem.event deliver` and `<S>.sem.theme medical-supplies`). For the next partial ASR result shown in the figure, the explicit subframe correctly adds several additional frame elements which formalize the meaning of the phrase *provide transportation* in this scenario as having the army move the clinic out of the market area.

To understand more quantitatively how well this technique works, we evaluated this technique in the SASO-EN test corpus, using different probability thresholds in the range [0.5,1.0). We present the results in Figure 4. To understand the effect of the threshold τ, note that, in general, the effect of selecting a higher threshold should be to "cherry pick" those frame elements which are most likely to appear in the complete frame G_u, thereby increasing precision while decreasing recall of the frame elements in S_j^{SUB} in relation to G_u. In the figure, we can see that this is indeed the case. The lines marked "(complete frame)"

[2]Note that no utterances in our *training* set were annotated.

Explicit subframe (with frame element probabilities)	**Predicted complete frame**	**Annotated subframe**
Partial ASR result: *we will*		
0.856 <S>.mood declarative 0.824 <S>.sem.agent captain-kirk 0.663 <S>.sem.modal.intention will 0.663 <S>.sem.speechact.type promise 0.776 <S>.sem.type event	<S>.mood declarative <S>.sem.agent captain-kirk <S>.sem.event deliver **<S>.sem.modal.intention will** **<S>.sem.speechact.type promise** <S>.sem.theme medical-supplies **<S>.sem.type event**	<S>.mood declarative **<S>.sem.modal.intention will** **<S>.sem.speechact.type promise**
Partial ASR result: *we will provide transportation*		
0.991 <S>.mood declarative 0.990 <S>.sem.agent captain-kirk 0.927 <S>.sem.event move 0.905 <S>.sem.instrument us-army 0.964 <S>.sem.modal.intention will 0.927 <S>.sem.source market 0.964 <S>.sem.speechact.type promise 0.928 <S>.sem.theme clinic 0.989 <S>.sem.type event	<S>.mood declarative **<S>.sem.agent captain-kirk** **<S>.sem.event move** **<S>.sem.instrument us-army** **<S>.sem.modal.intention will** **<S>.sem.source market** **<S>.sem.speechact.type promise** **<S>.sem.theme clinic** **<S>.sem.type event**	<S>.mood declarative **<S>.sem.agent captain-kirk** **<S>.sem.event move** **<S>.sem.instrument us-army** **<S>.sem.modal.intention will** **<S>.sem.source market** **<S>.sem.speechact.type promise** **<S>.sem.theme clinic** **<S>.sem.type event**

Figure 3: Explicit subframes and predicted complete frames for two partial ASR results in a user utterance of we *will provide transportation at no cost*.

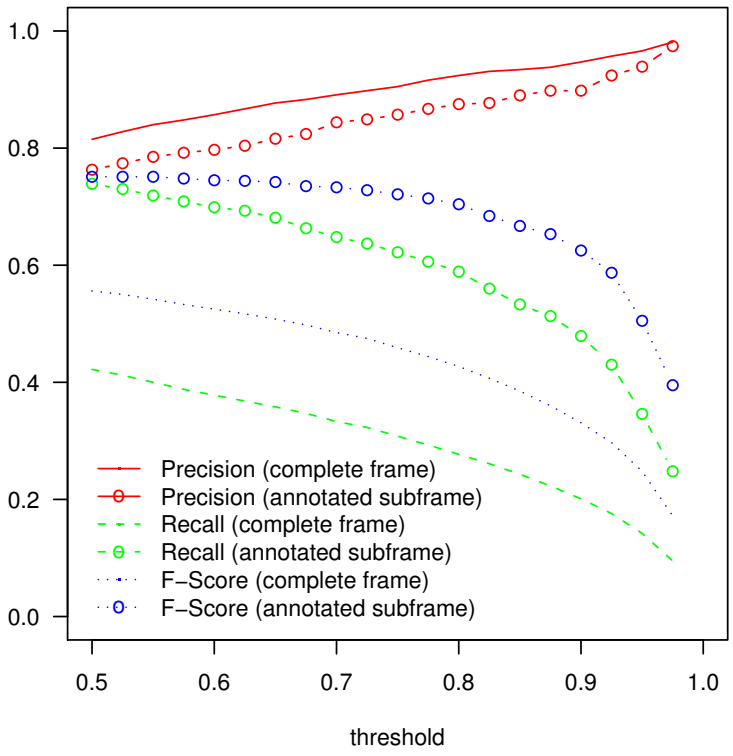

Figure 4: The effect of threshold on precision, recall, and F-Score of explicit subframes. All scores are measured in relation to complete utterance frames and annotated subframes.

in the figure evaluate the returned subframes in relation to the complete frame G_u associated with the user's *complete* utterance. We see that this method enables us to select subsets of frame elements that are most likely to appear in G_u: by increasing the threshold, it is possible to return subframes which are of increasingly higher precision in relation to the final frame G_u, but that also have lower recall.

We also evaluated the returned subframes in relation to the hand-annotated subframes, to assess its performance at identifying the user's explicit meaning. For an utterance u that generates partial ASR results $\langle r_1, ..., r_m \rangle$, we denote the hand-annotated subframe corresponding to partial ASR result r_j by G_j^{SUB}. In the lines marked "(annotated subframe)", we show the precision, recall, and F-score of the explicit subframe for each ASR result r_j in relation to the annotated subframe G_j^{SUB}.

As a first observation, note that at any threshold level, the explicit subframes do better at recalling the hand-annotated subframe elements than they do at recalling the complete frame elements. This means our new method is better at recalling what has been said already by the user than it is at predicting what will be said, as intended. We have seen two examples of this already, for the partial ASR result *hello* in Figure 2, and for the partial ASR result *we will* in Figure 3.

A second observation in Figure 4 is that precision remains better against the complete utterance frame than against the hand-annotated subframe (at all threshold levels). This indicates that the explicit subframes are often still predicting some aspects of the full frame. An example of this is given in Figure 5, where the user's partial utterance *we need to* is assigned an explicit subframe that includes frame elements describing an event of moving the clinic, which the user has not said explicitly. This happens because, in the SASO-EN domain, in fact there is nothing else that the interlocutors need to do besides move the clinic. So based on the NLU training data, the data-driven probabilities assigned by Equation (1) describe the additional frame elements as about as probable as the ones capturing the *we need to* part of the semantics (given at the right).

Finally, a third observation is that overall, the precision, recall, and F-score results against the annotated subframes using our method are surprisingly strong. For example, when evaluating the explicit subframes over all partial ASR results, an F-score of 0.75 is attained at thresholds in the range 0.5-0.55. This F-score is substantially better than the F-score of our predictive NLU in relation to the final full frames, which is 0.56 when evaluated over all partial ASR results. This means that our proposed model works better as an explicit incremental NLU than mxNLU works as a predictive incremental NLU. Further, we observe that this F-score of 0.75 against hand-annotated subframes is approximately as good as the F-score of 0.76 that is achieved when mxNLU is used to interpret complete utterances. We therefore conclude that the proposed model is a promising and viable approach to explicit incremental NLU in SASO-EN.

4 Discussion and Related Approaches

In this section, we discuss some of the practical aspects of using the technique presented here, in relation to some alternative approaches.

An important consideration for NLU techniques is the cost, in both time and knowledge, of the annotation that is needed. One attractive aspect of our technique is that the only semantic annotation that is required is the association of complete user utterances with complete NLU output frames. This task can be performed by anyone familiar with the scenario and the semantic frame format, such as a system developer or scenario designer. In fact, the annotation of the SASO-EN data set we use in this paper has been described in (DeVault et al., 2010), which reports that the overall corpus of 4678 token utterances was semantically annotated at an average rate of about 10 seconds per unique utterance.

The model in Equation (2) is what (Heintze et al., 2010) call a *hybrid output* approach, in which larger and larger frames are provided as partial input grows, but in which a detailed alignment between surface text and frames is not provided by the incremental NLU component. They contrast hybrid output systems with techniques that deliver either *whole-frame output* (like the predictive mxNLU) or *aligned output* that connects individual words to their meanings. A data-driven approach to providing aligned outputs would involve preparing a more detailed annotated corpus that aligns individual words and surface expressions to their corresponding frame elements. Given such a word-aligned corpus, one could train several kinds of models to produce the aligned outputs incrementally. One strategy would be to use a sequential tagging model such as a CRF to tag partial utterances with the frame elements that capture their explicit meaning, as in (Heintze et al., 2010).

Using a machine learning approach that models a more detailed alignment between surface text and frames would be one way to more cleanly separate explicit from predictive aspects of meaning. Preparing the training data for such models, however, would create additional annotation costs. As part of creating the annotated subframes for the evaluation presented in Section 3, we measured the time requirement for such annotation of word-meaning alignments at about 30 seconds per unique utterance. Performing full word-meaning alignment therefore takes about three times as much time as the complete utterance annotation needed for our technique. Ad-

Explicit subframe (with frame element probabilities)	**Predicted complete frame**	**Annotated subframe**
Partial ASR result: *we*		
0.753 <S>.mood declarative 0.687 <S>.sem.agent captain-kirk 0.692 <S>.sem.type event	<S>.mood declarative <S>.sem.agent captain-kirk <S>.sem.event deliver <S>.sem.modal.possibility can <S>.sem.speechact.type offer <S>.sem.theme medical-supplies <S>.sem.type event	
Partial ASR result: *we need to*		
0.945 <S>.mood declarative 0.928 <S>.sem.agent captain-kirk 0.900 <S>.sem.event move 0.816 <S>.sem.modal.deontic must 0.900 <S>.sem.source market 0.900 <S>.sem.speechact.type statement 0.906 <S>.sem.theme clinic 0.930 <S>.sem.type event	<S>.mood declarative <S>.sem.agent captain-kirk <S>.sem.event move <S>.sem.modal.deontic must <S>.sem.source market <S>.sem.speechact.type statement <S>.sem.theme clinic <S>.sem.type event	<S>.mood declarative <S>.sem.modal.deontic must <S>.sem.speechact.type statement

Figure 5: Explicit subframes and predicted complete frames for two partial ASR results in a user utterance of *we need to move the clinic*.

ditionally, this task requires a greater degree of linguistic knowledge and sophistication, as the annotator must be able to segment the utterance and align specific surface segments with potentially complex aspects of meaning such as modality, polarity, speech act types, and others. An example of the kinds of complexities that arise is illustrated in Figure 3, where the relationship between specific words like "provide" and "transportation" to frame elements like <S>.sem.event move and <S>.sem.theme clinic is not transparent, even if it is straightforward to mark the whole utterance as conveying that meaning in this domain. We have generally found this alignment task challenging for people without advanced linguistics training.

The reason we describe the method in this paper as an *approximation* of explicit NLU is that, partly because it is trained without detailed word-meaning alignments, it can be expected to occasionally include some predictive aspects of user utterance meaning. An example of this is the method's explicit subframe output for the phrase *we need to* in Figure 5.

Another way to approximate explicit NLU would be using the method (Heintze et al., 2010) call an *ensemble of classifiers*; it involves training an individual classifier for each frame key. Like the method presented here, an ensemble of classifiers can be easily trained to predict those frame elements that will appear in the *final* frame G_u for each utterance. And like our method, prediction with an ensemble of classifiers does not require detailed annotation of word-meaning alignment in the training data. One difference is that, with our method, by selecting an appropriate threshold, it is easy to enforce certain consistency properties on subframe outputs. In an ensemble of classifiers approach, there is no immediate guarantee that the output frame constructed by the independent classifiers will be internally consistent from the standpoint of downstream system modules (Heintze et al., 2010). For example, in the SASO-EN domain, an NLU frame should not contain frame elements that mix aspects of events and states in the SASO-EN ontology; e.g., the frame element <S>.sem.type event should not co-occur in an NLU output frame with the frame element <S>.sem.object-id market (which would be appropriate for a state frame but not for an event frame). With the method proposed here, if we select a threshold τ that is greater than 0.5, and if none of the complete NLU frames contain incompatible key values (which is relatively easy to enforce as part of the annotation task), then it will be mathematically impossible for two incompatible frame elements to be returned in a subframe.[3]

Ultimately, a classification method that is trained on word-meaning aligned data and that uses additional techniques to ensure that only valid, grammatical output frames are produced could prove to be an attractive approach. In future work, we will explore such techniques, and compare both their performance as well as their annotation and development costs to the approximation technique presented here.

5 Conclusion

The analysis in this paper has explored a method of approximating explicit incremental NLU using predictive

[3]Suppose frame element e_i is incompatible with e_j, and that $P(e_i \in G_u | r_j) > 0.5$. By stipulation, no complete frame $S \in \mathcal{S}$ such that $e_i \in S$ will also contain e_j. Since we know that the total probability of all the frames containing e_i must be greater than 0.5 in order for e_i to be selected, we can infer that the total probability of all frames including e_j must be less than 0.5, and thus that e_j will not be selected.

techniques in finite semantic domains. We have shown that an estimate of a user's explicit utterance meaning can be derived from an existing predictive understanding model in an example domain. We have quantified the performance of this new method in a corpus evaluation, showing that the method returns incremental explicit subframes with performance – as measured by precision, recall, and F-Score against hand-annotated subframes – that is competitive with a current statistical, data-driven approach for understanding complete spoken utterances in the same domain. We have provided examples that illustrate its strengths and weaknesses, and discussed the annotation costs associated with implementing this technique in relation to some alternative approaches. The method requires no additional annotation beyond what is needed for training an NLU module to understand complete spoken utterances. (Hand annotation of word-meaning alignment for a small number of utterances may be performed in order to tune the selected threshold and evaluate explicit understanding performance.) The method provides a free parameter that can be used to target the most advantageous levels of precision and recall for a particular dialogue system application. In future work, we will explore additional machine learning models that leverage richer training data, and investigate further the combination of explicit and predictive techniques.

Acknowledgments

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred. This material is based upon work supported by the National Science Foundation under Grant No. IIS-1219253. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Hendrik Buschmeier, Timo Baumann, Benjamin Dosch, Stefan Kopp, and David Schlangen. 2012. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 295–303, Seoul, South Korea, July. Association for Computational Linguistics.

Okko Buß and David Schlangen. 2011. Dium - an incremental dialogue manager that can produce self-corrections. In *Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.

David DeVault, Susan Robinson, and David Traum. 2010. IORelator: A graphical user interface to enable rapid semantic annotation for data-driven natural language understanding. In *Fifth Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation*.

David DeVault, Kenji Sagae, and David Traum. 2011a. Incremental interpretation and prediction of utterance meaning for interactive dialogue. *Dialogue & Discourse*, 2(1).

David DeVault, Kenji Sagae, and David R. Traum. 2011b. Detecting the status of a predictive incremental speech understanding model for real-time decision-making in a spoken dialogue system. In *Interspeech*, pages 1021–1024.

Arno Hartholt, Thomas Russ, David Traum, Eduard Hovy, and Susan Robinson. 2008. A common ground for virtual humans: Using an ontology in a natural language oriented virtual human architecture. In European Language Resources Association (ELRA), editor, *Proc. LREC*, Marrakech, Morocco, may.

Silvan Heintze, Timo Baumann, and David Schlangen. 2010. Comparing local and sequential models for statistical incremental natural language understanding. In *The 11th Annual Meeting of the Special Interest Group in Discourse and Dialogue (SIGDIAL 2010)*.

David Huggins-Daines, Mohit Kumar, Arthur Chan, Alan W. Black, Mosur Ravishankar, and Alex I. Rudnicky. 2006. Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In *Proceedings of ICASSP*.

Kenji Sagae, Gwen Christian, David DeVault, and David R. Traum. 2009. Towards natural language understanding of partial speech recognition results in dialogue systems. In *NAACL HLT*.

David Schlangen, Timo Baumann, and Michaela Atterer. 2009. Incremental reference resolution: The task, metrics for evaluation, and a bayesian filtering model that is sensitive to disfluencies. In *SIGDIAL*.

Ethan O. Selfridge, Iker Arizmendi, Peter A. Heeman, and Jason D. Williams. 2012. Integrating incremental speech recognition and pomdp-based dialogue systems. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 275–279, Seoul, South Korea, July. Association for Computational Linguistics.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of EACL 2009*.

David Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proc. of Intelligent Virtual Agents Conference IVA-2008*.

David Traum, David DeVault, Jina Lee, Zhiyang Wang, and Stacy C. Marsella. 2012. Incremental dialogue understanding and feedback for multi-party, multimodal conversation. In *The 12th International Conference on Intelligent Virtual Agents (IVA)*, Santa Cruz, CA, September.

David Traum. 2003. Semantics and pragmatics of questions and answers for dialogue agents. In *Proc. of the International Workshop on Computational Semantics*, pages 380–394, January.