Joint Identification and Segmentation of Domain-Specific Dialogue Acts for Conversational Dialogue Systems

Fabrizio Morbini and **Kenji Sagae**
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive, Playa Vista, CA 90094
{morbini,sagae}@ict.usc.edu

Abstract

Individual utterances often serve multiple communicative purposes in dialogue. We present a data-driven approach for identification of multiple dialogue acts in single utterances in the context of dialogue systems with limited training data. Our approach results in significantly increased understanding of user intent, compared to two strong baselines.

1 Introduction

Natural language understanding (NLU) at the level of speech acts for conversational dialogue systems can be performed with high accuracy in limited domains using data-driven techniques (Bender et al., 2003; Sagae et al., 2009; Gandhe et al., 2008, for example), provided that enough training material is available. For most systems that implement novel conversational scenarios, however, enough examples of user utterances, which can be annotated as NLU training data, only become available once several users have interacted with the system. This situation is typically addressed by bootstrapping from a relatively small set of hand-authored utterances that perform key dialogue acts in the scenario or from utterances collected from wizard-of-oz or role-play exercises, and having NLU accuracy increase over time as more users interact with the system and more utterances are annotated for NLU training.

While this can be effective in practice for utterances that perform only one of several possible system-specific dialogue acts (often several dozens), longer utterances that include multiple dialogue acts pose a greater challenge: the many available combinations of dialogue acts per utterance result in sparse coverage of the space of possibilities, unless a very large amount of data can be collected and annotated, which is often impractical. Users of the dialogue system, whose utterances are collected for further NLU improvement, tend to notice that portions of their longer utterances are ignored and that they are better understood when they express themselves with simpler sentences. This results in generation of data heavily skewed towards utterances that correspond to a single dialogue act, making it difficult to collect enough examples of utterances with multiple dialogue acts to improve NLU, which is precisely what would be needed to make users feel more comfortable with using longer utterances.

We address this chicken-and-egg problem with a data-driven NLU approach that segments and identifies multiple dialogue acts in single utterances, even when only short (single dialogue act) utterances are available for training. In contrast to previous approaches that assume the existence of enough training data for learning to segment utterances, e.g. (Stolcke and Shriberg, 1996), or to align specific words to parts of the formal representation, e.g. (Bender et al., 2003), our framework requires a relatively small dataset, which may not contain any utterances with multiple dialogue acts. This makes it possible to create new conversational dialogue system scenarios that allow and encourage users to express themselves with fewer restrictions, without an increased burden in the collection and annotation of NLU training data.

2 Method

Given (1) a predefined set of possible dialogue acts for a specific dialogue system, (2) a set of utterances

each annotated with a single dialogue act label, and (3) a classifier trained on this annotated utterance-label set, which assigns for a given word sequence a dialogue act label with a corresponding confidence score, our task is to find the best sequence of dialogue acts that covers a given input utterance. While short utterances are likely to be covered entirely by a single dialogue act that spans all of its words, longer utterances may be composed of spans that correspond to different dialogue acts.

bestDialogueActEndingAt($Text$,pos) **begin**
  if $pos < 0$ **then**
    return $\langle pos, \langle null, 1\rangle\rangle$;
  end
  $S = \{\}$;
  for $j = 0$ **to** pos **do**
    $\langle c, p\rangle$ = classify(words($Text, j, pos$));
    $S = S \cup \{\langle j, \langle c, p\rangle\rangle\}$;
  end
  return $\underset{\langle k,\langle c,p\rangle\rangle \in S}{\text{argmax}} \{p \cdot p' : \langle h, \langle c', p'\rangle\rangle =$ *bestDialogueActEndingAt*$(Text, k-1)\}$;
end

Algorithm 1: The function classify(T) calls the single dialogue act classifier subsystem on the input text T and returns the highest scoring dialogue act label c with its confidence score p. The function words(T, i, j) returns the string formed by concatenating the words in T from the i^{th} to the j^{th} included. To obtain the best segmentation of a given text, one has to work its way back from the end of the text: start by calling $\langle k, \langle c, p\rangle\rangle$ = *bestDialogueActEndingAt*$(Text, numWords)$, where $numWords$ is the number of words in $Text$. If $k > 0$ recursively call *bestDialogueActEndingAt*$(Text, k-1)$ to obtain the optimal dialogue act ending at $k-1$.

Algorithm 1 shows our approach for using a single dialogue act classifier to extract the sequence of dialogue acts with the highest overall score from a given utterance. The framework is independent of the particular subsystem used to select the dialogue act label for a given segment of text. The constraint is that this subsystem should return, for a given sequence of words, at least one dialogue act label and its confidence level in a normalized range that can be used for comparisons with subsequent runs. In the work reported in this paper, we use an existing data-driven NLU module (Sagae et al., 2009), developed for the SASO virtual human dialogue system (Traum et al., 2008b), but retrained using the data described in section 3. This NLU module performs maximum entropy multiclass classification, using features derived from the words in the input utterance, and using dialogue act labels as classes.

The basic idea is to find the best segmentation (that is, the one with the highest score) of the portion of the input text up to the i^{th} word. The base case S_i would be for $i = 1$ and it is the result of our classifier when the input is the single first word. For any other $i > 1$ we construct all word spans $T_{j,i}$ of the input text, containing the words from j to i, where $1 \leq j \leq i$, then we classify each of the $T_{j,i}$ and pick the best returned class (dialogue act label) $C_{j,i}$ (and associated score, which in the case of our maximum entropy classifier is the conditional probability $Score(C_{j,i}) = P(C_{j,i}|T_{j,i})$). Then we assign to the best segmentation ending at i, S_i, the label $C_{k,i}$ iff:

$$k = \underset{1 \leq h \leq i}{\text{argmax}} \left(Score(C_{h,i}) \cdot Score(S_{h-1})\right) \quad (1)$$

Algorithm 1 calls the classifier $O(n^2)$ where n is the number of words in the input text. Note that, as in the maximum entropy NLU of Bender et al. (2003), this search uses the "maximum approximation," and we do not normalize over all possible sequences. Therefore, our scores are not true probabilities, although they serve as a good approximation in the search for the best overall segmentation.

We experimented with two other variations of the argument of the argmax in equation 1: (1) instead of considering $Score(S_{h-1})$, consider only the last segment contained in S_{h-1}; and (2) instead of using the product of the scores of all segments, use the average score per segment: $(Score(C_{h,i}) \cdot Score(S_{h-1}))^{1/(1+N(S_{h-1}))}$ where $N(S_i)$ is the number of segments in S_i. These variants produce similar results; the results reported in the next section were obtained with the second variant.

3 Evaluation

3.1 Data

To evaluate our approach we used data collected from users of the TACQ (Traum et al., 2008a) dia-

logue system, as described by Artstein et al. (2009). Of the utterances in that dataset, about 30% are annotated with multiple dialogue acts. The annotation also contains for each dialogue act the corresponding segment of the input utterance.

The dataset contains a total of 1,579 utterances. Of these, 1,204 utterances contain only a single dialogue act, and 375 utterances contain multiple dialogue acts, according to manual dialogue act annotation. Within the set of utterances that contain multiple dialogue acts, the average number of dialogue acts per utterance is 2.3.

The dialogue act annotation scheme uses a total of 77 distinct labels, with each label corresponding to a domain-specific dialogue act, including some semantic information. Each of these 77 labels is composed at least of a core speech act type (e.g. wh-question, offer), and possibly also attributes that reflect semantics in the domain. For example, the dialogue act annotation for the utterance *What is the strange man's name?* would be `whq(obj: strangeMan, attr: name)`, reflecting that it is a wh-question, with a specific object and attribute. In the set of utterances with only one speech act, 70 of the possible 77 dialogue act labels are used. In the remaining utterances (which contain multiple speech acts per utterance), 59 unique dialogue act labels are used, including 7 that are not used in utterances with only a single dialogue act (these 7 labels are used in only 1% of those utterances). A total of 18 unique labels are used only in the set of utterances with one dialogue act (these labels are used in 5% of those utterances). Table 1 shows the frequency information for the five most common dialogue act labels in our dataset.

The average number of words in utterances with only a single dialogue act is 7.5 (with a maximum of 34, and minimum of 1), and the average length of utterances with multiple dialogue acts is 15.7 (maximum of 66, minimum of 2). To give a better idea of the dataset used here, we list below two examples of utterances in the dataset, and their dialogue act annotation. We add word indices as subscripts in the utterances for illustration purposes only, to facilitate identification of the word spans for each dialogue act. The annotation consists of a word interval and a dialogue act label[1].

Single DA Utt.	[%]	Multiple DA Utt.	[%]
Wh-questions	51	Wh-questions	31
Yes/No-questions	14	Offers to agent	24
Offers to agent	9	Yes answer	11
Yes answer	7	Yes/No-questions	8
Greeting	7	Thanks	7

Table 1: The frequency of the dialogue act classes most used in the TACQ dataset (Artstein et al., 2009). The left column reports the statistics for the set of utterances annotated with a single dialogue act the right those for the utterances annotated with multiple dialogue acts. Each dialogue act class typically contains several more specific dialogue acts that include domain-specific semantics (for example, there are 29 subtypes of wh-questions that can be performed in the domain, each with a separate domain-specific dialogue act label).

1. $\langle_0$ *his* $_1$ *name,* $_2$ *any* $_3$ *other* $_4$ *information* $_5$ *about* $_6$ *him,* $_7$ *where* $_8$ *he* $_9$ *lives* $_{10}\rangle$ is labeled with: [0 2] `whq(obj: strangeMan, attr: name)`, [2 7] `whq(obj: strangeMan)` and [7 10] `whq(obj: strangeMan, attr: location)`.

2. $\langle_0$ *I* $_1$ *can't* $_2$ *offer* $_3$ *you* $_4$ *money* $_5$ *but* $_6$ *I* $_7$ *can* $_8$ *offer* $_9$ *you* $_{10}$ *protection* $_{11}\rangle$ is labeled with: [0 5] `reject`, [5 11] `offer(safety)`.

3.2 Setup

In our experiments, we performed 10-fold cross-validation using the dataset described above. For the training folds, we use only utterances with a single dialogue act (utterances containing multiple dialogue acts are split into separate utterances), and the training procedure consists only of training a maximum entropy text classifier, which we use as our single dialogue act classifier subsystem.

For each evaluation fold we run the procedure described in Section 2, using the classifier obtained from the corresponding training fold. The segments present in the manual annotation are then aligned with the segments identified by our system (the

[1] Although the dialogue act labels could be thought of as compositional, since they include separate parts, we treat them as atomic labels.

alignment takes in consideration both the word span and the dialogue act label associated to each segment). The evaluation then considers as correct only the subset of dialogue acts identified automatically that were successfully aligned with the same dialogue act label in the gold-standard annotation.

We compared the performance of our proposed approach to two baselines; both use the same maximum entropy classifier used internally by our proposed approach.

1. The first baseline simply uses the single dialogue act label chosen by the maximum entropy classifier as the only dialogue act for each utterance. In other words, this baseline corresponds to the NLU developed for the SASO dialogue system (Traum et al., 2008b) by Sagae et al. (2009)[2]. This baseline is expected to have lower recall for those utterances that contain multiple dialogue acts, but potentially higher precision overall, since most utterances in the dataset contain only one dialogue act label.

2. For the second baseline, we treat multiple dialogue act detection as a set of binary classification tasks, one for each possible dialogue act label in the domain. We start from the same training data as above, and create N copies, where N is the number of unique dialogue acts labels in the training set. Each utterance-label pair in the original training set is now present in all N training sets. If in the original training set an utterance was labeled with the i^{th} dialogue act label, now it will be labeled as a positive example in the i^{th} training set and as a negative example in all other training sets. Binary classifiers for each N dialogue act labels are then trained. During run-time, each utterance is classified by all N models and the result is the subset of dialogue acts associated with the models that labeled the example as positive. This baseline is excepted to be much closer in performance to our approach, but it is incapable of determining what words in the utterance correspond to each dialogue act[3].

[2]We do not use the incremental processing version of the NLU described by Sagae et al., only the baseline NLU, which consist only of a maximum entropy classifier.

[3]This corresponds to the transformation of a multi-label classification problem into several binary classifiers, described as PT4 by Tsoumakas and Katakis (**?**).

		P [%]	R [%]	F [%]
Single	this	73	77	75
	$2^{nd}bl$	86	71	78
	$1^{st}bl$	82	77	80
Multiple	this	87	66	75
	$2^{nd}bl$	85	55	67
	$1^{st}bl$	91	39	55
Overall	this	78	72	75
	$2^{nd}bl$	86	64	73
	$1^{st}bl$	84	61	71

Table 2: Performance on the TACQ dataset obtained by our proposed approach (denoted by "this") and the two baseline methods. *Single* indicates the performance when tested only on utterances annotated with a single dialogue act. *Multiple* is for utterances annotated with more than one dialogue act, and *Overall* indicates the performance over the entire set. **P** stands for precision, **R** for recall, and **F** for F-score.

3.3 Results

Table 2 shows the performance of our approach and the two baselines. All measures show that the proposed approach has considerably improved performance for utterances that contain multiple dialogue acts, with only a small increase in the number of errors for the utterances containing only a single dialogue act. In fact, even though more than 70% of the utterances in the dataset contain only a single dialogue act, our approach for segmenting and identifying multiple dialogue acts increases *overall* F-score by about 4% when compared to the first baseline and by about 2% when compared to the second (strong) baseline, which suffers from the additional deficiency of not identifying what spans correspond to what dialogue acts. The differences in F-score over the entire dataset (shown in the *Overall* portion of Table 2) are statistically significant ($p < 0.05$). As a drawback of our approach, it is on average 25 times slower than our first baseline, which is incapable of identifying multiple dialogue acts in a utterance[4]. Our approach is still about 15% faster than our second baseline, which

[4]In our dataset, our method takes on average about 102ms to process an utterance that was originally labeled with multiple dialogue acts, and 12ms to process one annotated with a single dialogue act.

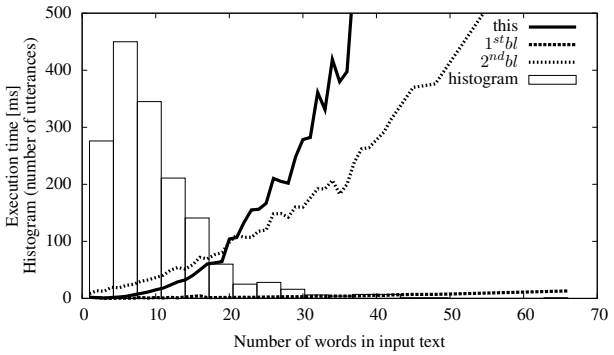

Figure 1: Execution time in milliseconds of the classifier with respect to the number of words in the input text.

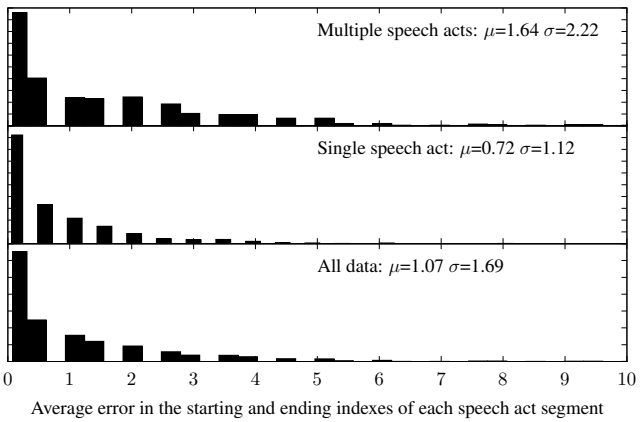

Figure 2: Histogram of the average absolute error in the two extremes (i.e. start and end) of segments corresponding to the dialogue acts identified in the dataset.

identifies multiple speech acts, but without segmentation, and with lower F-score. Figure 1 shows the execution time versus the length of the input text. It also shows a histogram of utterance lengths in the dataset, suggesting that our approach is suitable for most utterances in our dataset, but may be too slow for some of the longer utterances (with 30 words or more).

Figure 2 shows the histogram of the average error (absolute value of word offset) in the start and end of the dialogue act segmentation. Each dialogue act identified by Algorithm 1 is associated with a starting and ending index that corresponds to the portion of the input text that has been classified with the given dialogue act. During the evaluation, we find the best alignment between the manual annotation and the segmentation we computed. For each of the aligned pairs (i.e. extracted dialogue act and dialogue act present in the annotation) we compute the absolute error between the starting point of the extracted dialogue act and the starting point of the paired annotation. We do the same for the ending point and we average the two error figures. The result is binned to form the histogram displayed in figure 2. The figure also shows the average error and the standard deviation. The largest average error happens with the data annotated with multiple dialogue acts. In that case, the extracted segments have a starting and ending point that in average are misplaced by about ± 2 words.

4 Conclusion

We described a method to segment a given utterance into non-overlapping portions, each associated with a dialogue act. The method addresses the problem that, in development of new scenarios for conversational dialogue systems, there is typically not enough training data covering all or most configurations of how multiple dialogue acts appear in single utterances. Our approach requires only labeled utterances (or utterance segments) corresponding to a single dialogue act, which tends to be the easiest type of training data to author and to collect.

We performed an evaluation using existing data annotated with multiple dialogue acts for each utterance. We showed a significant improvement in overall performance compared to two strong baselines. The main drawback of the proposed approach is the complexity of the segment optimization that requires calling the dialogue act classifier $O(n^2)$ times with n representing the length of the input utterance. The benefit, however, is that having the ability to identify multiple dialogue acts in utterances takes us one step closer towards giving users more freedom to express themselves naturally with dialogue systems.

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. We would also like to thank the anonymous reviewers for their helpful comments.

References

Ron Artstein, Sudeep Gandhe, Michael Rushforth, and David R. Traum. 2009. Viability of a simple dialogue act scheme for a tactical questioning dialogue system. In *DiaHolmia 2009: Proceedings of the 13th Workshop on the Semantics and Pragmatics of Dialogue*, page 43–50, Stockholm, Sweden, June.

Oliver Bender, Klaus Macherey, Franz Josef Och, and Hermann Ney. 2003. Comparison of alignment templates and maximum entropy models for natural language understanding. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics - Volume 1*, EACL '03, pages 11–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sudeep Gandhe, David DeVault, Antonio Roque, Bilyana Martinovski, Ron Artstein, Anton Leuski, Jillian Gerten, and David R. Traum. 2008. From domain specification to virtual humans: An integrated approach to authoring tactical questioning characters. In *Proceedings of Interspeech*, Brisbane, Australia, September.

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*.

Andreas Stolcke and Elizabeth Shriberg. 1996. Automatic linguistic segmentation of conversational speech. In *Proc. ICSLP*, pages 1005–1008.

David R. Traum, Anton Leuski, Antonio Roque, Sudeep Gandhe, David DeVault, Jillian Gerten, Susan Robinson, and Bilyana Martinovski. 2008a. Natural language dialogue architectures for tactical questioning characters. In *Army Science Conference*, Florida, 12/2008.

David R. Traum, Stacy Marsella, Jonathan Gratch, Jina Lee, and Arno Hartholt. 2008b. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *IVA*, pages 117–130.