Classifying Dialog Acts in Human-Human and Human-Machine Spoken Conversations

Silvia Quarteroni and Giuseppe Riccardi

DISI - University of Trento
38050 Povo - Trento, Italy
{silviaq,riccardi}@disi.unitn.it

Abstract

Dialog acts represent the illocutionary aspect of the communication; depending on the nature of the dialog and its participants, different types of dialog act occur and an accurate classification of these is essential to support the understanding of human conversations. We learn effective discriminative dialog act classifiers by studying the most predictive classification features on Human-Human and Human-Machine corpora such as LUNA and SWITCHBOARD; additionally, we assess classifier robustness to speech errors. Our results exceed the state of the art on dialog act classification from reference transcriptions on SWITCHBOARD and allow us to reach a very satisfying performance on ASR transcriptions.

1. Introduction

Dialog Act (DA) classification is an important Spoken Language Understanding phase that consists of identifying the illocutionary acts of communication corresponding to an utterance. The above is a complementary process to concept extraction: indeed, the same concept may occur in a question, an answer or a clarification request. Different conversation types exist in terms of nature of speakers (human or machine), task-orientation and level of initiative, and while DA taxonomies such as DAMSL [1] have been applied to different dialog corpora (e.g. [2]), it can be noted that different conversation contexts lead to different DA distributions.

In this work, we select the most suitable features to classify DAs given a conversation type. In particular, we present: 1) a study on the efficiency of a number intra-utterance and extra-utterance features for DA classification on different corpora and 2) an analysis of different types of features and their suitability in different conversational domains and contexts. In this paper, we discuss related work in Section 2 and illustrate our machine learning models for DA classification in Section 3; Section 4 then describes our datasets and Sections 5-6 report our experiments both on manual transcriptions and on ASR data; finally, Section 7 summarizes our findings and future work.

2. Related work

Dialog Act classification has been applied to a number of corpora and DA taxonomies, including ICSI, AMI and SWITCHBOARD [2, 3, 4]. While varying greatly in terms of granularity and domain-dependence, most taxonomies include the three main groups of *core* DAs, expressing the basic communication functions such as informing, asking and answering, *feedback*, representing clarification requests and acknowledgments, and *conventional*, such as greeting and thanking [1].

In this work, we focus on DA classification based on textual features, so we do not consider prosodic and acoustic features as in e.g. [5]. While word n-grams are agreed to be vital features in DA classification [2, 3, 4], some studies argue against the use of extra-utterance features such as previous dialog acts [3].

In terms of learning models, both discriminative and generative approaches have been deployed in the classification task: for instance, [2] achieves 71% accuracy on on a fixed train/test SWITCHBOARD split by using a generative approach based on word trigrams. In this paper, we opt for discriminative classifiers and particularly Support Vector Machines (SVMs), found effective as standalone or in accordance with other methods in a number of DA classification studies [5, 4].

3. Learning Dialog Act classifiers

In order to classify dialog acts using a supervised learning method, a labelled dialog corpus must be available. In our model, a dialog is represented as an ordered list of turns t_i, each bearing an utterance u_i pronounced by a speaker. The latter is an ordered list of dialog acts da_{ij}, characterized by their index j within the utterance, their label l_{ij} (chosen from a given taxonomy) and surface (i.e. word-level realization) s_{ij}. Given such a representation, the learning process is articulated into two main phases: feature extraction from the data and classifier learning.

In the former, given a dialog act da_{ij}, we study the contribution of the following features to determine l_{ij}:

UNI, BI, TRI word unigrams, bigrams, and trigrams in s_{ij};

LEN number of words in s_{ij};

PIN the label of the DA immediately preceding da_{ij} in u_i;

1ST first word in s_{ij};

TID turn index i;

POUT label of the last DA in t_{i-1}, if any.

We can regard the first three features as intra-utterance features and the last two as extra-utterance features. The latter are particularly interesting as by nature, DA classification differs by the presence of backward- and forward- looking dialog acts (see [1]) from other classification tasks where the user utterance can be regarded as "self-contained", such as question [6] or intent [7] classification. For instance, *"I'd like to rent a car"* could be a statement or an answer depending on the context.

Furthermore, it can be noted that PIN and POUT are dynamic features, i.e. not available "offline". In our current experiments, we use the value of PIN and POUT as found in the reference annotation; indeed, related work has either ignored

This work was partly funded by the European Commission project ADAMACH (contract 022593).

such features or shown that dynamically acquired DA labels yield lower classification results than when ignoring such features [8]. In the learning phase, we use SVMs, effective and robust textual data classifiers provided that a sufficient number of training examples is available; in particular, our classifiers are learned using a linear kernel SVM implemented by CMU's MinorThird [9].

4. Data

We work with two datasets, LUNA [10] and SWITCHBOARD [2]; these differ by dialog act taxonomy, task, interaction type, language and size, as described in the following subsections.

4.1. LUNA

The LUNA corpus is composed of hardware/software troubleshooting dialogs in Italian; here, dialog acts have been annotated according to the ADAMACH taxonomy [10], a compact version of DAMSL consisting of 16 DA classes: core (*info-request, action-request, yes-answer, no-answer, answer, offer, report-action, inform*), conventional (*greet, quit, apology, thank*), and feedback (*clarif-request, ack, filler*), with the addition of *other* and *non-interpretable*.

The first of LUNA's two subsets, LUNA-HM, was acquired with a Wizard of Oz approach (WOZ): the wizard reacts to user's spontaneous spoken requests belonging to one of ten possible dialog scenarios inspired by the services provided by an Italian customer care company. LUNA-HM contains 174 dialogs annotated with 2112 caller dialog acts; as an example, the utterance: *"Hi, my printer isn't working this morning"* contains a *greet* DA followed by *inform*. Out of the caller DAs, 1753 have been randomly chosen for training and 359 for testing.

The second subset, LUNA-HH, is a Human-Human corpus where dialogs refer to real user conversations engaged in a software/hardware troubleshooting task; it contains 94 annotated dialogs from 2 speakers: the caller and the service provider. Out of the 2352 caller dialog acts, 1983 have been randomly chosen for training and 370 for testing. It is striking to note the difference in the number of dialog acts callers pronounce when they think they are interacting with a machine (WOZ dialogs) and when they are talking to a human (HH dialogs). Indeed, "real" dialogs contain in average an almost doubled number of caller dialog acts when compared to the WOZ ones.

An analysis of DA distribution in the official training and test splits of LUNA-HM and LUNA-HH, reported in Table 1, reveals that the most frequent DAs in HM dialog are answers (*answer, yes-answer*), showing a low degree of user initiative; in contrast, HH dialogs show a higher frequency of *inform*, feedback moves such as *ack* and even conversational fillers.

4.2. SWITCHBOARD

SWITCHBOARD [2] consists of 1155 human-human dialogs where speakers were free to converse on a given topic; hence, dialogs are not task-oriented, although the topic of the conversation was fixed beforehand. Dialog act annotation followed the 42-class compact DAMSL taxonomy [1]. In the experiments reported in [2], 19 of the 1155 dialogs have been retained for testing and 1115 for training; the latter form partitions $sw00$ to $sw10$ of the official release [11]. Table 2 reports the most frequent dialog acts in the different partitions of the corpus used in our experiments.

Table 1: Top caller dialog acts by frequency in the training and test subsets of the LUNA-HM and LUNA-HH corpora

LUNA-HM			LUNA-HH		
Dialog act	%train	%test	Dialog act	%train	%test
answer	33%	35%	inform	24%	23%
y-ans	14%	17%	ack	23%	32%
thank	8%	8%	filler	8%	5%
ack	7%	7%	answer	7%	9%
inform	7%	7%	info-req	7%	6%
filler	7%	7%	yes-ans	7%	9%
greet	5%	6%	other	7%	5%
no-ans	3%	3%	thank	3%	2%
quit	3%	3%	quit	3%	3%
info-req	2%	3%	rep-act	2%	2%
clarif-req	2%	2%	clarif-req	2%	1%
other	2%	3%	offer	2%	1%
TOTAL	1753	359	TOTAL	1983	370

Table 2: Dialog acts ranked by frequency in SWITCHBOARD: partitions $sw00$ and $sw01$, full dataset and official test set

Dialog act	% $sw00$	% $sw01$	% $full$	% $test$
Statement-non-op	36%	43%	36%	31%
Ack	18%	16%	19%	18%
Statement-op	17%	10%	13%	17%
Uninterpr	8%	7%	6%	9%
Agree	5%	6%	5%	5%
YN-question	2%	2%	2%	2%
Non-verb	2%	2%	2%	2%
Apprec	1%	1%	2%	2%
Y-Ans	1%	1%	1%	2%
Conv-closing	$<$1%	1%	1%	2%
Wh-Question	1%	1%	1%	1%
Remaining DAs	9%	10%	12%	12%

5. Experiments

We report experiments on both LUNA and SWITCHBOARD.

5.1. LUNA

Our results on LUNA-HM (Table 3) show that while the unigram baseline model achieves an ER of 22.7% on the official train/test split, the best performing model we have found achieves 17% by combining the following features: unigrams, bigrams, the first word, the length of the instance the previous turn's last dialog act and the latest dialog act in the current turn (UNI+BI+1ST+LEN+PIN+POUT). When analyzing class-by-class error rates (Figure 4), we note first that the first word helps most in *yes-answer*, while LEN helps in *answer* and *filler*; secondly, bigrams and furthermore the final feature combination are useful for *inform*. Finally, it can be noted that the best model drastically reduces the ER on *answer* and *inform* with respect to the other models: this is certainly due to the presence of both PIN and POUT, providing useful context to distinguish between these two classes.

LUNA-HH results (Table 3) show that classifying HH conversation is much more challenging than the HM case: indeed, the unigram model baseline yields a 44.6% ER. We believe this is due to several factors, one of which is the degree of initiative in which humans engage in a conversation with another human. This translates into a much larger lexicon with respect to HM conversation: indeed, while LUNA-HM contains about 1400 words, LUNA-HH contains 2100.

Table 3: LUNA classification results

LUNA-HM		LUNA-HH	
Model	ER	Model	ER
UNI	22.7%	UNI	44.6%
+BI	21.9%	+BI	46.5%
+BI+LEN	21.6%	+PIN	45.1%
+BI+POUT	21.6%	+LEN	44.3%
+BI+PIN	21.0%	+1ST	43.8%
+BI+PIN+POUT	20.5%	+POUT	37.8%
+BI+1ST	20.5%	+LEN+POUT	37.8%
+BI+TID	19.9%	+1ST +POUT	37.8%
+BI+LEN+PIN+POUT	19.3%		
+BI+1ST+LEN+PIN+POUT	17.0%		

Table 4: Classification ER of the most frequent LUNA classes

	ER on LUNA-HM (%)						
Model	All	ans	y-ans	thk	info	fil	ack
UNI	22.7	15.7	6.6	0.0	57.1	27.3	69.6
+BI	21.9	16.5	6.6	0.0	**50.0**	22.7	73.9
+BI+POUT	21.6	14.2	8.2	3.3	57.1	22.7	78.3
+BI+LEN	21.6	**11.0**	13.1	0.0	57.1	**18.2**	73.9
+BI+1ST	20.5	14.2	**4.9**	3.3	57.1	27.3	69.6
+BI+LEN+1ST +PIN+POUT	17.1	4.7	13.1	3.3	**21.4**	27.3	69.6

	ER on LUNA-HH (%)						
Model	All	info	ack	fil	answer	y-ans	oth
UNI	44.6	33.7	18.8	64.7	82.4	81.3	88.9
+LEN	44.3	33.7	18.8	64.7	79.4	**78.1**	**83.3**
+1ST	43.8	30.2	17.9	64.7	79.4	84.4	88.9
+POUT	37.8	**27.9**	20.5	**58.8**	**64.7**	**40.6**	88.9
+LEN+1ST +PIN+POUT	36.8	**25.6**	19.7	**58.8**	55.9	34.4	88.9

On the LUNA-HH corpus, the best performing model reduces the ER to 37.8% by combining word unigrams with the previous turn's last dialog act. The lower efficiency of the bigram feature can be explained by the data sparsity problem affecting LUNA-HH: indeed, the fact that this corpus contains more words than LUNA-HM translates into twice as many bigrams: while the former contains about 6K bigrams, LUNA-HH contains 11K. Another interesting point is that the POUT feature, representing the latest act from the other speaker, reveals to be sufficient to reach the lowest ER. Also notably, while the ER of *ack* is around 70% in LUNA-HM, it is below 20% in LUNA-HH: this is probably because *ack* is much more frequent in the LUNA-HH training set than in the LUNA-HM set (Tab. 1). By examining class-by-class results (Table 4), we note the following: first, LEN helps most in *other, yes-answer*, DAs whose surface realization is generally short; moreover, POUT helps most in *answer, filler, yes-answer, inform*: indeed the previous DA is vital to discriminate between *answer* and *inform*, or *yes-answer* from a generic *ack*; finally, the combination of all features is beneficial for the highly confusable *inform* and *answer*.

5.2. SWITCHBOARD

As SWITCHBOARD is a very large corpus and the official testset accounts for only 19 of the 1155 dialogs [2], we first study the most effective classification features in a 5-fold cross-validation regime on two of its 13 partitions, $sw00$ and $sw01$ (100 dialogs each, comparable to the LUNA corpora). As shown in Table 2 (columns 2-3), the most frequent DA in $sw00$ and $sw01$ is *Statement-non-opinion*, followed by *Acknowledgement* then *Statement-opinion*.

Table 5: SWITCHBOARD classification results: partitions $sw00$ and $sw01$ in 5-fold cross-validation (xval) and on the official testset ($test$), full training set ($full$) on $test$

	$sw00$		$sw01$		$full$
Model	xval	$test$	xval	$test$	$test$
UNI	29.0 ± 0.5	32.9	27.6 ± 0.1	37.2	30.0
+BI	28.9 ± 0.9	32.6	27.7 ± 0.7	35.4	30.0
+BI+LEN	27.7 ± 0.9	31.6	26.9 ± 0.5	34.8	-
+BI+LEN+PIN	26.2 ± 0.6	29.5	24.3 ± 0.2	32.7	27.6
Stolcke et al. [2]					29.0

As illustrated in Table 5, in the $sw00$ and $sw00$ datasets, the bigram feature is effective; this can be explained by the presence of a sufficient number of dialogs to discriminate between features. Another interesting observation is that in this corpus, the PIN feature is particularly effective; indeed, SWITCHBOARD contains more DAs per turn, hence several DAs appear between the current DA and the latest DA of the previous turn (POUT). In addition, utterances are often composed of a sequence of DAs with the same label, and PIN captures this. In both $sw00$ and $sw01$ the best performace was achieved by combining unigrams, bigrams, instance length and the latest dialog act found in the current turn (Table 5). Furthermore, it can be observed in Table 5 (col. $sw01$ - xval) that our cross-validation experiments on subset $sw01$, a portion of the SWITCHBOARD corpus as small as 100 dialogs, achieve an ER of 24.3% with our best model and 27.7% with a simple bigram model. Indeed, [12] also obtained a lower ER[1] than [2] when choosing a test set formed by 173 randomly selected dialogs instead of the "official" one.

Having identified the most interesting features in a cross-validation regime, we have experimented with the same features by training on $sw00$ and $sw01$ and testing on the official test set[2]. This allowed to single out the best feature combination, that resulted to be the same combination obtained during the cross-validation experiment in both cases. In particular, when analyzing the ER on the most frequent $sw00$ classes (Table 6), we note that: a) LEN mostly helps to classify *Statement-non opinion* and *Agree*; b) PIN mostly helps to classify *Agree* and *Uninterpretable*; c) the combined model is beneficial to further reduce the error on *Statement-opinion* and *Statement-non opinion* (as in LUNA where *inform* and *answer* were distinguished).

Eventually, after finding the most effective feature combinations on different subsets of the SWITCHBOARD corpus, we tried them on the full corpus by training and testing on the official splits. Our results, reported in Table 5 (col. $full$), show that the best performing model found during feature selection reduces the state-of-the-art ER of 29% [2] by 5% relative. Table 6 (bottom) illustrates the error rates of different learning models on the most frequent DA classes when training on $full$ and testing on $test$: with the exception of *Statement-non-opinion* the ER decreases of 15-20% relative on each class when applying the best model.

Finally, Figure 1 illustrates the learning curve of the SWITCHBOARD Dialog Act classifier for increasing amounts

[1] ER of 23.8% using intra-utterance features, e.g. n-grams and supertags, and further improvement when adding extra-utterance features

[2] Now available at: `www.stanford.edu/~jurafsky/ws97/`

Table 6: ER on SWITCHBOARD official testset when training on $sw00$ and $full$ (overall and most frequent classes)

	ER of $sw00$ on $test$ (%)					
Model	All	St-nn-op	Ack	St-op	Unint	Agr
UNI	32.9	20.2	3.9	46.9	30.4	68.5
+ BI	32.6	19.9	4.7	48.0	30.4	68.5
+ BI+ LEN	31.6	**17.7**	4.9	46.9	31.2	**60.7**
+ BI+ PIN	30.0	18.3	5.4	43.1	**23.4**	**52.9**
+ BI+LEN+PIN	29.5	**17.1**	5.3	**42.0**	23.1	52.9
	ER of $full$ on $test$ (%)					
Model	Overall	St-nn-op	Ack	St-op	Unint	Agr
UNI	30.0	13.2	5.2	54.2	25.6	61.7
+ BI+LEN+PIN	27.6	17.9	3.9	44.2	21.4	52.9

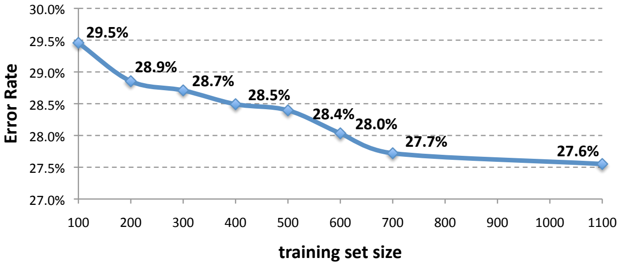

Figure 1: SWITCHBOARD ER for increasing training set size

of training data. We have gradually increased the training set starting from 100 dialogs (the first partition of the training set, i.e. $sw00$) to the complete training set size. Our first interesting finding is that the first two partitions (200 dialogs composing partitions $sw00$ – $sw01$) are sufficient to reach and slightly outperform the 29% ER achieved by [2]; however, there is a margin for improvement left as around 700 dialogs are sufficient to reduce the ER to a value close to that obtained when training on the full training set. The drop in ER achieved between these two "boundaries" provides an interesting insight over the potentials of DA classification in the LUNA domain; while in the latter the classification problem is simplified by the presence of fewer classes, both datasets contain less than 200 dialogs.

6. Experiments on ASR transcriptions

The greatest challenge of DA classification in Spoken Dialog Systems is dealing with ASR transcriptions, which require utterance segmentation in DAs and robustness with respect to recognition error. In our initial experiments, we have computed the drop in accuracy when dealing with ASR on LUNA-HM turns that contain only 1 DA. We create our models based on the manually transcribed LUNA-HM training split and the best feature combination as found in our previous experiments (Sec. 5.1), and compare performances when testing on manual versus ASR transcriptions of the 154 test set utterances containing 1 DA. As illustrated in Table 7, the relative drop in performance on LUNA-HM is about 14% when testing on the ASR transcription instead of the manual one. This is comparable to the 13% relative drop obtained for SWITCHBOARD in [2] when working with the ASR top hypothesis instead of the manually transcribed text. Interestingly, on the latter corpus Word Error Rate was around 40%, while in LUNA-HM it is "only" 27%; we argue that the smaller size of the LUNA dataset plays a role.

Table 7: Classification results on LUNA ASR transcriptions

Dataset	Model	ER (%)
LUNA-HM (TRS)	UNI+BI+1ST+LEN+PIN+POUT	11.7%
LUNA-HM (ASR)	UNI+BI+1ST+LEN+PIN+POUT	13.6%

7. Conclusions

We have investigated the most effective features for Dialog Act (DA) classification on Human-Machine and Human-Human conversation corpora. Our results suggest that although DAs are generally domain-independent, the nature of conversation has an impact on the choice of classification features. In particular, the human or machine nature of speakers, their degree of initiative, and the dialog objective (task-oriented vs conversation) affect the distribution of dialog acts and the lexicon used. As a general conclusion, our methods are robust and meet or exceed state-of-the-art performance; in contrast to e.g. [3], we find that using extra-utterance features (e.g. previous DA) increases classifier accuracy, hence not only "local" features are effective. We are currently researching a discriminative model joining turn segmentation and DA classification in order to analyze conversations deriving directly from ASR.

Acknowledgement We thank Andreas Stolcke and Dan Jurafsky for releasing the SWITCHBOARD train/test splits used in [2].

8. References

[1] M. G. Core and J. F. Allen, "Coding dialogs with the DAMSL annotation scheme," in *Proc. AAAI Fall Symposium on Communicative Actions in Humans and Machines*, 1997.

[2] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, 2000.

[3] N. Webb, M. Hepple, and Y. Wilks, "Dialogue act classification based on intra-utterance features," in *Proc. AAAI Workshop on Spoken Language Understanding*, 2005.

[4] D. Verbree, R. Rienks, and D. Heylen, "Dialogue-act tagging using smart feature selection: results on multiple corpora," in *Proc. SLT*, Palm Beach, 2006.

[5] R. Fernandez and R. W. Picard, "Dialog act classification from prosodic features using support vector machines," *Speech Prosody*, 2002.

[6] A. Moschitti, S. Quarteroni, R. Basili, and S. Manandhar, "Exploiting syntactic and shallow semantic kernels for question/answer classification," in *Proc. ACL*, 2007.

[7] H. Alshawi, "Effective utterance classification with unsupervised phonotactic models," in *Proc. NAACL*, 2003.

[8] S. Bangalore, G. Di Fabbrizio, and A. Stent, "Learning the structure of task-driven human-human dialogs," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 7, 2008.

[9] W. W. Cohen, "Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data," 2004.

[10] M. Dinarelli, S. Quarteroni, S. Tonelli, A. Moschitti, and G. Riccardi, "Annotating spoken dialogs: from speech segments to dialog acts and frame semantics," in *Proc. SRSL*, 2009.

[11] J. J. Godfrey and E. Holliman, "Switchboard-1 release 2," Philadelphia, USA, 1997.

[12] S. Bangalore, G. Di Fabbrizio, and A. Stent, "Learning the structure of task-driven human-human dialogs," in *Proc. ACL*, 2006.