SEMANTIC ANNOTATIONS FOR CONVERSATIONAL SPEECH: FROM SPEECH TRANSCRIPTIONS TO PREDICATE ARGUMENT STRUCTURES

Arianna Bisazza, Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, Giuseppe Riccardi

DISI - University of Trento - 38050 Povo - Trento, Italy
{bisazza, dinarelli, silviaq, moschitti, riccardi}@disi.unitn.it, satonelli@fbk.eu

ABSTRACT

In this paper, we describe the semantic content, which can be automatically generated, for the design of advanced dialog systems. Since the latter will be based on machine learning approaches, we created training data by annotating a corpus with the needed content. Given a sentence of our transcribed corpus, domain concepts and other linguistic levels ranging from basic ones, i.e. part-of-speech tagging and constituent chunking level, to more advanced ones, i.e. syntactic and predicate argument structure (PAS) levels are annotated. In particular, the proposed PAS and taxonomy of dialog acts appear to be promising for the design of more complex dialog systems. Statistics about our semantic annotation are reported.

1. INTRODUCTION

Spoken language understanding (SLU) addresses the problem of extracting and annotating the meaning structure from spoken utterances in the context of human dialogs [6]. In spoken dialog systems (SDS) most used models of SLU are based on the identification of slots (entities) within one or more frames (frame-slot semantics) that is defined by the application. While this model is simple and clearly insufficient to cope with interpretation and reasoning, it has supported the first generation of spoken dialog systems. Such dialog systems are thus limited by the ability to parse semantic features such as predicates and to perform logical computation in the context of a specific dialog act [2]. Such limitation is reflected in the type of human-machine interactions which are mostly directed at querying the user for specific slots (e.g. 'What is the departure city?") or implementing simple dialog acts (e.g. confirmation). We believe that an important step in overcoming such limitation relies on the study of models of human-human dialogs at different levels of representations: lexical, syntactic, semantic and discourse. In this paper we present our results in addressing these issues in the context of the LUNA research project for next-generation spoken dialog interfaces [6]. We propose models for different levels of annotation of the SDS corpus including attribute-value, predicate argument structures and dialog acts. We describe the tools and the adaptation of off-the-shelf resources to carry out annotation of the predicate argument structures of spoken utterances. We present a quantitative analysis of such semantic structures for both human-machine and human-human conversations. To the best of our knowledge this is the first SDS corpus (human-machine and human-human) annotated with a multilayer approach to the annotation of lexical, semantic and dialog features. This allowed us to investigate statistical relations between the language processing layers such as shallow semantics and dialog strategies used by humans or machines. In the following sections we describe the product of our effort, i.e. the LUNA spoken dialog corpus, a quantitative analysis of the corpus and statistical correlations between annotation layers.

2. ANNOTATION MODEL IN LUNA

In the context of the European project LUNA[1] we are acquiring a corpus to study new solutions for Spoken Dialog Systems. The corpus will contain 1000 equally partitioned Human-Human (HH) and Human-Machine (HM) dialogs. These are recorded by CSI, an Italian customer care and technical support center. HH dialogs refer to real user conversations engaged in a problem solving task in the domain of software/hardware repairing. HM dialogs are 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 from the services provided by CSI. The above data is organized in transcriptions and annotations of speech based on a new multi-level protocol studied specifically within the LUNA project, i.e. the annotation levels of words, turns[2], dialog acts, attribute-values, predicate argument structures. The annotation at word level is made with part-of-speech and morphosyntactic information following the recommendations of EAGLES corpora annotation [7]. The attribute-value annotation uses a predefined domain ontology to specify concepts and their relations. Dialog acts are

This work was partially funded by the European Commission - LUNA project contract n. 33549.

[1] EU FP6 contract No. 33549

[2] A turn is defined as the interval when a speaker is active, between two pauses in his/her speech flow.

used to annotate intention in an utterance and can be useful to find relations between different utterances as the next section

Table 1. The ADAMACH dialog act taxonomy

Group	Dialog act tags
Core	Info-request, Action-request, Yes-answer, No-answer, Answer, Offer, ReportOnAction, Inform
Conventional	Greet, Quit, Apology, Thank
Feedback	ClarificationRequest, Ack, Filler
Non interpretable	Other

will show. For predicate structure annotation, we followed the FrameNet model [1] (see Section 2.2).

2.1. Dialog Act annotation

According to speech act theory [10], when pronouncing a sequence of words a speaker is either performing an action or trying to change the information state of the addressee. Dialog act annotation therefore consists in detecting and labeling the main function or goal of an utterance.

We performed such annotation using the ADAMACH dialog act taxonomy, designed following taxonomies such as [5] and [11], summarized in Table 1.

Dialog act annotation was performed manually by one annotator on speech transcriptions previously segmented into turns as mentioned above. The annotation unit for dialog acts is the utterance; however, utterances are complex semantic entities that do not necessarily correspond to turns. Hence, a segmentation of the dialog transcription into utterances was performed by the annotator before dialog act labeling. Both utterance segmentation and dialog act labeling were performed through the MMAX tool [8].

The annotator proceeded according to the following guidelines: 1) by default, a turn is also an utterance; 2) if more than one tag is applicable to an utterance, choose the tag corresponding to its main function; 3) in case of doubt among several tags, give priority to tags in *core* dialog acts group; 4) when needed, split the turn into several utterances or merge several turns into one utterance.

Utterance segmentation provides the basis not only for dialog act labeling but also for the other semantic annotations. See Fig. 1 for a dialog sample where each line represents an utterance annotated on the three levels.

2.2. Predicate Argument annotation

For the predicate-argument structure annotation layer, we adopted the FrameNet paradigm [1]. In this model, the meaning of predicates (or *lexical units*, usually verbs, nouns, or adjectives) is conveyed by *frames*, conceptual structures describing prototypical situations or events and the involved participants. Semantic roles (or *frame elements*) are the salient entities in the evoked situation and correspond to the syntactic dependents of the lexical units. They can be either core, i.e. typical of a given frame, non-core or peripheral, with several

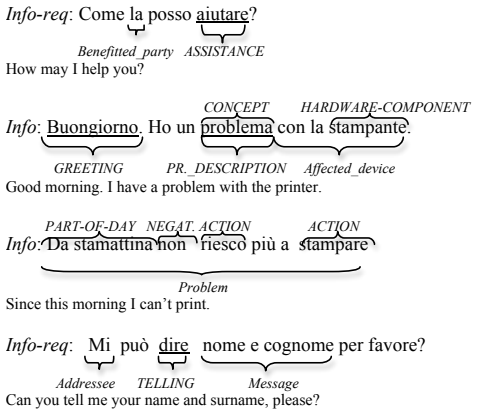


Fig. 1. Annotated dialog extract. Each utterance is preceded by dialog act annotation. Domain attribute annotation appears above the text, PAS annotation below the text.

instantiations in different frames and more generic meaning. All lexical units evoking the same frame have similar semantics and are attested with the same frame elements.

The FrameNet paradigm has been applied to develop the English FrameNet database, where annotated sentences from the British National Corpus and other smaller corpora provide evidence to frames and roles description. Since previous work (see e.g. [9]) has shown that the English FrameNet ontology is meaningful for other languages, we adopted the language-independent part of the FrameNet database and we instantiated its frames and roles definition on 50 HM and 50 HH dialogs from the Italian LUNA corpus (figures will increase in the near future). Figure 1 shows a dialog sample with PAS annotation reported below the utterance. All lexical units are underlined and the frame is written in capitals, while the other labels refer to frame elements. In particular, *ASSISTANCE* is evoked by the lexical unit *aiutare* and has one attested frame element (*Benefitted_party*), while *GREETING* has no frame element and *PROBLEM_DESCRIPTION* and *TELLING* two frame elements each.

Figure 2 gives a comprehensive view of the annotation process, from audio file transcription to the annotation of three semantic layers. Whereas domain attribute and PAS annotation are carried out on the segmented dialogs at utterance level, PAS annotation requires POS-tagging and syntactic parsing (via Bikel's parser trained for Italian[4]), because frame information points to parse-tree nodes.

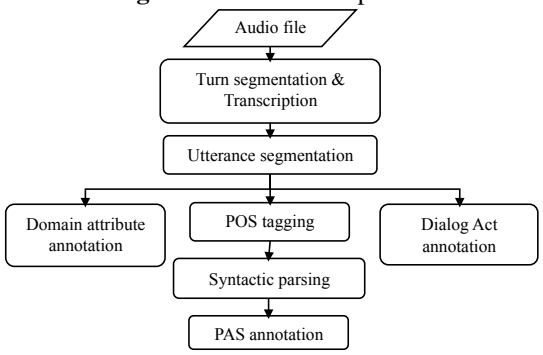
3. EVALUATION OF THE ANNOTATION

We evaluated the outcome of dialog act and PAS annotation levels on both the HH and HM corpora by not only analyzing frequencies and occurrences in the separate levels, but also their interaction, as discussed in the following sections.

3.1. Dialog Act annotation

Analyzing the annotation available so far for 50 HM and 50 HH dialogs on the dialog act level, we note that in average an

Fig. 2. The annotation process



HH dialog is composed of 48.9±17.4 (Std. Dev.) dialog acts, whereas a HM dialog is composed of 18.9±4.4. The difference between average lengths shows how HH spontaneuous speech can be redundant, while HM dialogs are more limited to an exchange of essential information. The standard deviation of a conversation in terms of dialog acts is considerably higher in the HH corpus than in the HM one. This can be explained by the fact that the WOZ follows a unique, previously defined task-solving strategy that does not allow for digressions. Utterance segmentation was also performed differently on the two corpora. In HH we performed 167 turn mergings and 225 turn splittings; in HM dialogs no turn merging was performed, but only turn splittings (158).

Table 2 reports the dialog acts occurring in the HM and HH corpora ranked by their frequencies. From a comparative analysis, we notice that: 1) *info-request* is by far the most common dialog act in HM, whereas in HH *ack* and *info* share the top ranking position; 2) the most frequently occurring dialog act in HH, i.e. *ack*, is only ranked 11th in HM; 3) *clarification-request*'s relative frequency (4,7%) is considerably higher in HH than in HM.

We also analyzed the ranking of the most frequent dialog act bigrams in the two corpora. We can summarize our comparative analysis to the following: in both corpora, most bigram types contain *info* and *info-request*, as expected in a troubleshooting system. However, the bigram *info-request answer*, which we expected to form the core of a task-solving dialog, is only ranked 5th in the HH corpus, while 5 out of the top 10 bigram types contain *ack*. We believe that this is because HH dialogs primarily contain spontaneous information-providing turns (e.g. several *info info* by the same speaker) and acknowledgements for the purpose of backchannel. Instead, HM dialogs, structured as sequences of *info-request answers* pairs, are more minimal and brittle, showing how users tend to avoid redundancy when addressing a machine.

3.2. Predicate Argument annotation

We annotated an initial set of 50 HM and 50 HH dialogs with frame information using the *Salto* tool [3]. We identified all lexical units corresponding to *semantically relevant* verbs, nouns and adjectives with a syntactic subcategorization pattern. In particular, we annotated all lexical units that imply an action, introduce the speaker's opinion or describe the office environment. We decided to adopt the original FrameNet description of frames and frame elements, introducing new frames and roles only in case of gaps in the FrameNet ontology. In particular, we introduced 20 new frames out of the 174 taken from FrameNet because the original definition of frames related to hardware / software, data-handling and customer assistance was too coarse-grained. Few new frame elements were introduced as well, mostly expressing syntactic realizations that are typical of spoken Italian.

Table 2. Dialog acts ranked by frequency in the two corpora

human-machine (HM)			human-human (HH)		
DA	cnt	rel.freq.	DA	cnt	rel.freq.
info-req	249	26,3%	ack	582	23,8%
answer	171	18,1%	info	562	23,0%
info	163	17,2%	info-req	303	12,4%
y-ans	70	7,4%	answer	192	7,8%
quit	60	6,3%	clarif	116	4,7%
thank	56	5,9%	offer	114	4,7%
greet	50	5,3%	y-ans	112	4,6%
offer	49	5,2%	quit	101	4,1%
clarif	26	2,7%	rep-act	91	3,7%
act-req	25	2,6%	other	70	2,9%
ack	12	1,3%	act-req	69	2,8%
filler	6	0,6%	filler	61	2,5%
n-ans	5	0,5%	thank	33	1,3%
other, rep-act	2	0,2%	n-ans	26	1,1%
apol	1	0,1%	greet, apol	7	0,3%
TOTAL	947		TOTAL	2446	

Table 3 shows some statistics about the corpus dimension and the average annotated information. HH dialogs show less frame instances in average than HM, meaning that speech disfluencies, that are not present in turns uttered by the machine, negatively affect the semantic density of a turn. The standard deviation w.r.t. frames is higher in HH dialogs than in the HM ones. Similarly to DA annotation, it depends on the task-solving strategy of HM dialogs, while HH dialogs are richer in digressions.

Table 4 reports the 10 most frequent frames occurring in the HM and HH dialogs. The relative frame frequency in HH dialogs is sparser than in HM ones, meaning that the turns uttered by the machine influence the discourse topic and that the semantics of HH dialogs is more variable. The most frequent frame group comprises frames related to information exchange that is typical of the help-desk activity, including *Telling*, *Greeting*, *Contacting*, *Statement*, *Recording*, *Communication*. Another relevant group encompasses frames related to the operational state of a device, for example *Being_operational*, *Change_operational_state*, *Operational_testing*, *Being_in_operation*.

Table 3. Corpus statistics

	HM	HH
Total number of turns	662	1997
Average dialog length in turns	13.2	39.9
Average turn length in tokens	11.4	10.8
Frame instances per dialog	18.5±5.1	39.0±17.2
Frame instances per turn	1.4	1.0
Frame elements per fr. inst.	1.6	1.7

3.3. Mutual information between PAS and dialog acts

A unique feature of the LUNA corpus is the availability of both a semantic and a dialog act annotation level: it is intuitive to seek relationships in the purpose of improving the recognition and understanding of each level by using features from the other. We considered a subset of 20 HH and 50 HM dialogs and computed an initial analysis of the co-occurrences of dialog acts and PAS. We noticed that each PAS tended to co-occur only with a limited subset of the available dialog act tags, and moreover in most cases the co-occurrence happened with only one dialog act. For a more thorough analysis, we computed the weighted mutual information between PAS and dialog acts.

In the HM corpus, we noted some interesting associations between dialog acts and PAS. First, `info-req` has the maximal MI with PAS like *Being_in_operation* and *Being_attached*, as requests are typically used by the operator to get information about the status of device. Several PAS denote a high MI with the `info` dialog act, including *Activity_resume, Information, Being_named, Contacting*, and *Resolve_problem*. As for the remaining acts, `clarif` has the highest MI with *Perception_experience* and *Statement*, used to warn the addressee about understanding problems resp. asking him to repeat/rephrase an utterance. The `answer` tag is highly informative with PAS referring to the exchange of information (*Read_data*) or to actions performed by the user after a suggestion of the system (*Change_operational_state*).

In the HH corpus, most of the PAS are highly mutually informative with `info`: indeed, as shown in Table 2, this is the most frequently occurring act in HH except for `ack`, which rarely contain verbs that can be annotated by a frame. As for the remaining acts, there is an easily explainable high MI between `quit` and *Greeting*; moreover, `info-req` denote its highest MI with *Giving*, as in requests to give information.Our MI results corroborate our initial observation that for most PAS, the mutual information tends to be very high in correspondence of one dialog act type: this suggests the beneficial effect of including shallow semantic information such as PAS as features for dialog act classification.

4. CONCLUSIONS

In this paper we have proposed a comprehensive framework for annotating human dialogs with lexical, semantic and discourse features. Such effort is crucial to investigate the complex dependencies between the layers of semantic processing. We have designed the annotation model to incorporate features and models developed both in the speech and language research community and bridging the gap between the two communities. Such multi-layer annotation corpus allows us the investigation of cross-layer dependencies and across human-machine and human-human dialogs as well as training of semantic models which accounts for the predicate interpretation.

Table 4. The 10 most frequent frames (* =newly introduced)

HM			HH		
Frame	cnt	freq-%	Frame	cnt	freq-%
Greeting*	146	15.8	Telling	143	7.3
Telling	134	14.5	Greeting*	124	6.3
Recording	83	8.9	Awareness	74	3.8
Being_named	74	8.0	Contacting	63	3.2
Contacting	52	5.6	Giving	62	3.2
Usefulness	50	5.4	Navigation*	61	3.1
Being_oper.	28	3.0	Chg._op._state	51	2.6
Probl._desc.*	24	2.6	Percept._exp.	46	2.3
Inspecting	24	2.6	Insert_data*	46	2.3
Percept._exp.	21	2.3	Come_to_sight*	38	1.9

5. REFERENCES

[1] C. F. Baker, C. J. Fillmore, and J. B. Lowe. The Berkeley FrameNet Project. In *ACL*, 1998.

[2] F. Bechet, G. Riccardi, and D. Hakkani-Tur. Mining spoken dialogue corpora for system evaluation and modeling. In *EMNLP*, 2004.

[3] A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal. Salto - a versatile multi-level annotation tool. In *LREC*, 2006.

[4] A. Corazza, A. Lavelli, and G. Satta. Analisi sintattica-statistica basata su costituenti. *Intelligenza Artificiale*, (2):38–39, 2007.

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

[6] R. De Mori, F. Bechet, D. Hakkani-Tur, M. McTear, G. Riccardi, and G. Tur. Spoken language understanding: A survey. In *IEEE Signal Processing Magazine*, 2008.

[7] G. Leech and A. Wilson. EAGLES recommendations for the morphosyntactic annotation of corpora. Technical report, ILC-CNR, 2006.

[8] C. Müller and M. Strube. Multi-level annotation in MMAX. In *SIGdial*, 2003.

[9] S. Padó. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. PhD thesis, Univ. Saarlandes, 2007.

[10] J. M. Sinclair and R. M. Coulthard. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. Oxford University Press, 1975.

[11] D. Traum. Dialogue management in conversational agency: The TRAINS-93 dialogue manager. In *TWLT*, 1996.