SIMULTANEOUS DIALOG ACT SEGMENTATION AND CLASSIFICATION FROM HUMAN-HUMAN SPOKEN CONVERSATIONS

Silvia Quarteroni, Alexei V. Ivanov, Giuseppe Riccardi

DISI - University of Trento
38050 Povo (Trento), Italy
quarteroni@elet.polimi.it, {ivanov|riccardi}@disi.unitn.it

ABSTRACT

An accurate identification dialog acts (DAs), which represent the illocutionary aspect of communication, is essential to support the understanding of human conversations. This requires 1) the segmentation of human-human dialogs into turns, 2) the intra-turn segmentation into DA boundaries and 3) the classification of each segment according to a DA tag. This process is particularly challenging when both segmentation and tagging are automated and utterance hypotheses derive from the erroneous results of ASR. In this paper, we use Conditional Random Fields to learn models for simultaneous segmentation and labeling of DAs from whole human-human spoken dialogs. We identify the best performing lexical feature combinations on the LUNA and SWITCHBOARD human-human dialog corpora and compare performances to those of discriminative D classifiers based on manually segmented utterances. Additionally, we assess our models' robustness to recognition errors, showing that DA identification is robust in the presence of high word error rates.

Index Terms— Dialog Acts, Spoken Language Understanding, Conditional Random Fields

1. INTRODUCTION

In Spoken Language Understanding, the identification of Dialog Acts (DAs) within an utterance, i.e. its illocutionary acts of communication, is a complementary process to concept extraction. Indeed, as the same concept may occur in a question, an answer or a clarification request, both levels of analysis are necessary for the complete understanding of conversations. Identifying DAs within an utterance is not a trivial task, as utterances may contain more than one DA; hence, prior to DA "classification", utterances must be segmented according to DA boundaries. In addition, the word error rates of current Automatic Speech Recognition (ASR) systems result in imperfect utterance hypotheses.

In this work, we use a discriminative approach, namely Conditional Random Fields (CRFs), to *simultaneously* segment an utterance into its DA boundaries and label such segments according to a DA tag. We experiment with different feature combinations and the well-known SWITCHBOARD corpus, reaching performances close to those of DA tagging alone. Moreover, we study the impact of using ASR as opposed to manual transcriptions on the LUNA Italian dialog corpus.

2. RELATED WORK

Traditionally, the problem of identifying the different DA segments within an utterance has been approached in a separate fashion: first, DA boundary segmentation within an utterance was addressed with generative or discriminative approaches [1, 2]; then, DA labels were assigned to such boundaries based on multi-classification [3, 4].

Work on utterance segmentation into DA boundaries includes [1], where a boosting approach combined weak learners; CRFs have been found to perform better than HMMs and maximum entropy approaches in [2] for the same task. Once DA segments have been identified, tagging them according to their DA tag becomes a multi-way classification problem. In [3], combinations of word n-grams and prosodic features were deployed in a semi-supervised learning setting to assign a unique DA label to an utterance, assuming that the latter contained a single DA. In previous work on DA classification [4], we have experimented with a SVM-based approach which achieved state-of-the-art results on the SWITCHBOARD reference transcriptions and set the baseline for the same task on the LUNA corpus.

In contrast to the above approaches, this work deals with *simultaneous* DA segmentation and tagging, a process that performs 1) segmentation of human-human dialogs into turns 2) intra-turn DA segmentation and 3) classification of each segment according to a DA tag in a single step. We propose CRFs as a learning model for their successful use in similar tasks, e.g. information extraction and shallow parsing [5]. Our work may be compared to [6], where a single perceptron is used for the same task. In this paper, we focus on class-by-class performances of different feature combinations, validat-

Work partly funded by EC project ADAMACH (contract 022593).

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

ing them on the exact SWITCHBOARD train/test split used in [7] for DA tagging, as well as on both manual and ASR transcriptions on the LUNA corpus [8].

3. LEARNING MODEL

In our learning model, a dialog is represented as an ordered list of turns t_i, each bearing an utterance u_i. The latter is annotated with an ordered list of dialog acts (DAs) da_{i0},..,da_{iN}, where DA values (l_{ij}) are taken from a DA taxonomy. Each DA da_{ij} is transcribed with a word sequence s_{ij}.

For instance, turn t_0, bearing utterance u_0: "Hi, my printer isn't working this morning", may be represented as the sequence of da_{00}, with label $l_{00} =$*greet* and surface $s_{00} =$ "Hi", and da_{01}, with value $l_{01} =$ *inform* and word sequence $s_{01} =$"my printer isn't working this morning".

Given such a representation, we formalize DA segmentation and classification as a sequence classification problem, i.e. the problem of predicting a single DA label l_{ij} that applies to a word sequence s_{ij} in a turn t_i (turn boundaries are not necessarily known). To estimate $P(l_{ij}|s_{ij})$ given our training dialogs, we use a combination of features extracted from the utterance: these mainly consist of lexical features such as word and Part-of-Speech (POS) n-grams (Section 5). As a learning algorithm, we use first-order linear-chain CRFs, a category of probabilistic learners frequently used for labeling and segmenting structured data [5]. CRFs are undirected graphical models used to specify the conditional probability of assigning output labels given a set of input observations. A conditional probability distribution is defined over label sequences given a particular observation sequence (of e.g. DA surfaces), rather than a joint distribution over both label and observation sequences. CRFs simultaneously segment and assign labels to the tokens of an unsegmented, unlabelled input.

4. CORPORA

We work with two datasets, SWITCHBOARD and LUNA, which differ by DA taxonomy, task, language and size.

SWITCHBOARD [7] consists of 1155 human-human dialogs in English; these are not task-oriented but topic-based conversational dialogs. DA annotation followed the 42-class compact DAMSL taxonomy [9]. In the DA classification experiments reported in [7], 19 of the 1155 dialogs have been retained for testing and 1115 for training; we use the same split in Section 5 to allow a direct comparison.

The LUNA corpus [8] consists of hardware/software troubleshooting dialogs in Italian; in particular, its LUNA-HH subset contains human-human conversations following ten dialog scenarios relating to the services provided by an Italian customer care company. LUNA-HH contains 94 dialogs from 2 speakers (the caller and the service provider), where utterances have been manually transcribed and annotated according to the DAs in the ADAMACH taxonomy, a compact version of DAMSL consisting of 16 DA classes [8] (see Table 1). Dialogs have randomly been split into a training, development and test set: while the training set contains 63 dialogs and 4686 DAs, the test set contains 15 dialogs and 742 DAs.

An analysis of DA distributions in the two sets, reported in Table 1, reveals that the most frequent DAs are spontaneous informative acts and acknowledgments, while questions (*info-req*) and answers together constitute less than 20% of the whole distribution. This illustrates the complexity of interaction and the value of identifying DAs rather than assuming pure question/answer interaction – even in strictly task-oriented dialog.

Table 1. Dialog act distribution in the LUNA-HH train/testset

LUNA-HH					
DA label	% train	% test	DA label	% train	% test
ack	22.4%	24.5%	offer	4.9%	3.9%
inform	22.0%	24.1%	quit	3.5%	3.6%
info-req	10.6%	10.5%	clarif-req	4.3%	2.9%
answer	5.5%	6.5%	act-req	2.3%	2.7%
other	6.3%	5.2%	thank	2.3%	1.3%
y-answer	4.7%	4.9%	no-answer	1.0%	0.7%
report-action	2.6%	4.5%	greet	0.2%	0.5%
filler	7.1%	3.9%	apology	0.3%	0.4%

5. EXPERIMENTS

We report experiments conducted on both SWITCHBOARD and LUNA-HH using the crfsuite[1] CRF implementation.

5.1. SWITCHBOARD

In our SWITCHBOARD experiments, we evaluated several models combining word and POS features[2]. We evaluated simultaneous DA segmentation and tagging with these models on the SWITCHBOARD test set using word-level and turn-level accuracy (Acc_W resp. Acc_T); these are respectively defined as the number of correct word-level predictions out of the total number of words and the number of correctly segmented and labelled turns out of the total number of turns.

As a baseline feature combination, we adopted word unigrams, in particular the words appearing in a range of [-2,..,+2] words with respect to the current word; this model gave 65% Acc_W and 51.5% Acc_T (Table 2, row UNIW). Subsequently, we augmented the UNIW features with bigrams containing the current word: this model (row WORDS) reached 68.2% Acc_W, as bigrams allow to account for the relative position of words in e.g. question vs statement form. Furthermore, we incremented the latter model with POS unigrams and bigrams in the [-2,..,+2] word range; the addition

[1] available at: www.chokkan.org/software/crfsuite

[2] POS tags were obtained automatically via the qtag state-of-the-art probabilistic tagger available at: phrasys.net/uob/om/software

of POS tags as a generalization method allows for a gain up to 69.3% in accuracy (row UNIBIPOS). Finally, we adopted the word & POS features as in the template of the CoNLL 2000 text chunking task[3]: this model (Table 2, row CoNLL) is equivalent to the WORDS model with the addition of POS unigrams, bigrams and trigrams in the [-2,..,+2] word range. With the latter, accuracy increases to 70.3%. Table 2 also reports DA alignment error rates[4] (DA ER), showing that the "CoNLL" model achieves 34.1% DA ER. In comparison, the SVM model in [4] achieved 72.5% DA *classification* accuracy (i.e. 27.5% DA ER) on the same data; automatic segmentation introduces a 24% relative ER increase.

Different feature combinations, such as widening the range of the n-gram features, did not yield further improvement. The good performance of the "CoNLL" model can be explained by noting that simultaneous DA segmentation and tagging can be viewed as the task of "chunking" an utterance to identify and label its segments' illocutionary roles.

Table 2. SWITCHBOARD: accuracy and DA ER of simultaneous segmentation and tagging models - reference transcriptions

Segm. & Tagging Model	Acc_W	Acc_T	DA ER
UNIW	65.0%	51.5%	40.5%
WORDS	68.2%	55.6%	52.2%
UNIBIPOS	69.3%	56.7%	35.2%
CoNLL	**70.9%**	**58.8%**	**34.1%**
CoNLL+POLAR	70.3%	58.4%	60.1%

When augmenting the granularity of the simultaneous segmentation and tagging task to *entire dialogs*, the CoNLL model only mildly deteriorates to 62.9% Acc_W, i.e. a relative decrease around 10% with respect to Acc_W with turns as input (Table 3, col. 2). Moreover, when using the CoNLL model and features for DA boundary segmentation only, we note that the segmentation Acc_W reaches a nearly optimal value on the SWITCHBOARD testset (98.6%), thanks to the large amount of training data and of the binary nature of the task: the same data is now used to only discriminate between boundary and non-boundary words (Table 3, col. 1).

Finally, simple segmentation yields the same Acc_W for turn and dialog input granularity. We believe the robustness of these results to input granularity are mainly due to the large amount of data available to train our models; moreover, extra-utterance information may be useful to distinguish specific dialog acts appearing at the beginning or end of a turn.

When analyzing the most frequent confusion pairs generated by our model, we note that out of the top five pairs, 70% involve opinion vs non-opinion statements, while the remaining ones affect agreements and acknowledgements. To account for this, we added the current word's prior polarity, as extracted from the OpinionFinder lexicon[5] to our best feature combination in order to connotate the current word in our algorithm. While this helped to reduce the confusability of opinion vs non-opinion statements, the overall accuracy did not improve (Table 2, row CoNLL+POLAR) due to the increase of other types of confusion pairs, such as *acknowledge* vs *statement-non-opinion*. We believe that polarity features deserve a deeper future study to be useful to this task.

Table 3. SWITCHBOARD: segmentation and tagging Acc_W for turn and dialog input granularity - reference transcriptions

Granularity	Segmentation Acc_W	Segm. & Tagging Acc_W
Turn	98.6% (CoNLL)	70.9% (CoNLL)
Dialog	98.6% (CoNLL) [majority: 85.7%]	62.9% (CoNLL)

5.2. LUNA-HH

On the LUNA-HH corpus, the baseline word unigram feature (UNIW) gave a 45.3% Acc_W, while we found the most effective feature to be the UNIBI feature combination (see Sec. 5.1), which gave a 48.8% Acc_W and a 25% Acc_T. POS features, automatically obtained via the Italian TreeTagger[6], did not improve the accuracy of our model: indeed, the best performing combination for SWITCHBOARD, CoNLL, gave 41.2% Acc_W. This is probably due to the small amount of data available for feature extraction, as suggested by our segmentation-only results below.

Table 4. LUNA-HH: segmentation and tagging Acc_W

	Reference		ASR
Granularity	Segmentation Acc_W	Segm. & Tagging Acc_W	
Turn	98.2% (CoNLL)	47.8% (UNIBI)	40.0% (UNIBI)
Dialog	89.9% (CoNLL) [majority: 82.3%]	47.6% (UNIBI)	43.0% (UNIBI)

When full dialogs are used as input, Acc_W decreases only slightly to 47.6% (Table 4), indicating that the greatest loss comes from simultaneous segmentation and tagging versus tagging alone, rather than from input granularity. Pure segmentation yields a similar picture: a 98.2% Acc_W for turn-level input, 89.9% for entire dialogs (Table 4). The best performing segmentation model is the feature-rich CoNLL model; indeed, as boundary detection is a binary classification problem, the greater number of instances per class reduces data sparseness and enables more features to contribute.

[3] www.cnts.ua.ac.be/conll2000/chunking/

[4] Elaborated by sclite, www.itl.nist.gov/iad/mig/tools/

[5] Within the OpinionFinder lexicon (www.cs.pitt.edu/mpqa/), each word is assigned a prior polarity denoting its property to carry an either positive or negative polarity regardless of a specific context

[6] available at: www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

5.3. 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 errors. For this reason, we validated the models learned for LUNA on the output of real ASR transcription of test set dialogs, characterized by 43% average WER.

Our results under such conditions were surprising: at turn level, DA segmentation and tagging gave a 40% Acc_W, thus relative degradation with respect to reference transcription was around 18% only. A similar relative degradation of 15% was observed by [7] for the SWITCHBOARD DA classification task, where ASR transcriptions had 40% WER. These findings suggest that a statistical method for segmenting and classifying DAs is effective at ignoring the word-level noise and goes to the core of the dialog semantics.

When performing segmentation and tagging using entire dialogs transcribed by ASR as input, word accuracy only slightly decreased to 43%. This may seem surprising as the accuracy at turn level under the same conditions is 40%; however, an attentive analysis suggests that the presence of dialog acts surrounding the current turn adds useful information and partly compensates ASR noise, making classification simpler.

Finally, we devised an experiment simulating realistic ASR word errors on the reference transcriptions. Each point of Figure 1 illustrates the average Acc_W of the UNIBI model for five versions the LUNA-HH testset, obtained by simulating ASR at a given WER between 0% and 50%. The simulation was done by randomly drawing from the probability distribution of recognized tokens (as estimated from real ASR data), conditioned on the corresponding reference tokens. Although such technique does not take into account mutual dependency between the adjacent errors, it appears to be reasonable approximation (compare results of Figure 1 and Table 4). As visible from Figure 1, accuracy loss is less steep than WER increase, confirming the robustness of the segmentation and tagging algorithm to ASR errors.

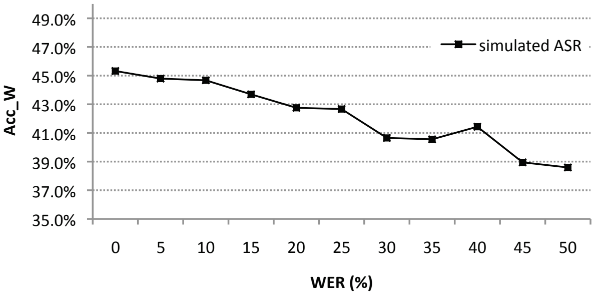

Fig. 1. LUNA-HH: Simultaneous segmentation and tagging Acc_W on simulated ASR transcriptions with different WER

6. CONCLUSIONS

We address the task of simultaneous dialog act segmentation and classification in human-human conversations via a discriminative approach based on Conditional Random Fields and explore the contribution of a number of lexical feature combinations. Our experiments, conducted over manual transcriptions of the SWITCHBOARD corpus and on both manual and ASR transcriptions of the Italian LUNA corpus, show that automatic segmentation implies an additional loss in accuracy of around 20% when compared to classification alone [4]. Despite this, our models for simultaneously segmenting and tagging of dialog acts exceed 70% accuracy on the SWITCHBOARD corpus. On LUNA, where we also compare the impact of ASR transcription to manual transcription, we find that our models are robust even in the presence of WER above 40%.

7. ACKNOWLEDGMENTS

This work was partially supported by the European Commission Marie Curie Excellence Grant for the ADAMACH project (contract No. 022593) and by the Livememories Project funded by Autonomous Province of Trento.

8. REFERENCES

[1] N.K. Gupta and S. Bangalore, "Segmenting Spoken Language Utterances into Claused for Semantic Classification," in *Proc. ASRU*. Citeseer, 2003, pp. 525–530.

[2] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech," in *Proc. ACL*, 2005.

[3] U. Guz, S. Cuendet, D. Hakkani-Tur, and G. Tur, "Multi-view semi-supervised learning for dialog act segmentation of speech," *IEEE TASLP*, vol. 18, no. 2, 2010.

[4] S. Quarteroni and G. Riccardi, "Dialog Act Classification in Human Human and Human Machine Conversations," in *Proc. INTERSPEECH*, 2010.

[5] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. ICML*, 2001.

[6] R. Granell, S. Pulman, C. Martinez-Hinarejos, and J. M. Benedi, "Dialog Act Tagging and Segmentation with a Single Perceptron," in *Proc. INTERSPEECH*, 2010.

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

[8] 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.

[9] 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.