Simultaneous dialogue act segmentation and labelling using lexical and syntactic features

Ramon Granell, Stephen Pulman
Oxford University Computing Laboratory,
Wolfson Building, Parks Road,
Oxford, OX1 3QD, England
ramg@comlab.ox.ac.uk
sgp@clg.ox.ac.uk

Carlos-D. Martínez-Hinarejos
Instituto Tecnológico de Informática,
Universidad Politécnica de Valencia,
Camino de Vera, s/n, 46022, Valencia, Spain
cmartine@dsic.upv.es

Abstract

Segmentation of utterances and annotation as dialogue acts can be helpful for several modules of dialogue systems. In this work, we study a statistical machine learning model to perform these tasks simultaneously using lexical features and incorporating deterministic syntactic restrictions. There is a slight improvement in both segmentation and labelling due to these restrictions.

1 Introduction

Dialogue acts (DA) are linguistic abstractions that are commonly accepted and employed by the the dialogue community. In the framework of dialogue systems, they can be helpful to identify and model user intentions and system answers by the dialogue manager. Furthermore, in other dialogue modules such as the automatic speech recognizer or speech synthesiser, DA information may be also used to increase their performance.

Many researchers have studied automatic DA labelling using different techniques. However, in most of this work it is common to assume that the dialogue turns are already segmented into separate utterances, where each utterance corresponds to just one DA label, as in (Stolcke et al (2000); Ji and Bilmes (2005); Webb et al (2005)). This is not a realistic situation because the segmentation of turns into utterances is not a trivial problem.

There have been many previous approaches to segmentation of turns prior to DA labelling, beginning with (Stolcke and Shriberg (1996)). Typically some combination of words and part of speech (POS) tags is used to predict segmentation boundaries. In this work we make use of a statistical model to solve both the DA labelling task and the segmentation task simultaneously, following (Ang et al (2005); Martínez-Hinarejos et al (2006)). Our aim is to see whether going beyond the word n-gram models can improve accuracy, using syntactic information (constituent structure) obtained from the dialogue transcriptions. We examine whether this information can improve the segmentation of the dialogue turns into DA segments. Intuitively, it seems logical to believe that most of these segments must coincide with particular syntactic structures, and that segment boundaries would respect constituent boundaries.

2 Dialogue data

The dialogue corpus used to perform the experiments is the Switchboard database (SWBD). It consists of human-human conversations by telephone about generic topics. There are 1155 5-minute conversations, comprising approximately 205000 utterances and 1.4 million words. The size of the vocabulary is approximately 22000 words.

All this corpus has been manually annotated at the dialogue act level using the SWBD-DAMSL scheme, (Jurafsky et al (1997)), consisting of 42 different labels. Every dialogue turn was manually segmented into utterances. The average number of segments (utterances) per dialogue turn is 1.78 with a standard deviation of 1.41. Each utterance was assigned one SWBD-DAMSL label (see Figure 1).

3 Syntactic analysis of DA segments

An initial analysis of the syntactic structures of the dialogue data was performed to study their possible relevance for DA segmentation.

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 333–336,
Queen Mary University of London, September 2009. ©2009 Association for Computational Linguistics

- $LAUGH he waits until it gets about seventeen below up here $SEG and then he calls us , $SEG
sd sd
- he waits until it gets about seventeen below up here and then he calls us .

Figure 1: The first row is an original segmented dialogue turn, where the $SEG label indicates the end of a DA segment. The second row contains the corresponding DA label for each segment, where ”sd” corresponds to the SWBD-DAMSL label of Statement non-opinion. The third row is the input for the parser.

3.1 Parsing of spontaneous dialogues

One of the main problems we face when we try to syntactically analyse a corpus transcribed from spontaneous speech by different people such as SWBD corpus, is the inconsistency of annotation conventions for spontaneous speech phenomena and punctuation marks. This can be problematic for parsers, as they work at the sentence level. Some of the dialogue turns of the SWBD corpus are not transcribed using consistent punctuation conventions. We therefore carried out some preprocessing so that all turns end with proper punctuation marks. Additionally, the non-verbal labels (e.g. $LAUGH, $OVERLAP, $SEG, ...) are removed. In Figure 1 there is an example of this process.

The Stanford Parser, (Klein and Manning (2003)) was used for the syntactic analysis of the transcriptions of SWBD dialogues. The English grammar used to train the parser is based on the standard LDC Penn Treebank WSJ training sections 2-21. Is is important to remark that the nature of the training corpus (journalistic style reports) is different from the transcriptions of spontaneous speech conversations. We would therefore expect a decrease in accuracy. As output of the parsing process, a tree that contains syntactic structures was provided (e.g. see Figure 2).

```
(ROOT
 (S (: -)
  (S
   (NP (PRP he))
    (VP (VBZ waits)
     (SBAR (IN until)
      (S
       (NP (PRP it))
       (VP (VBZ gets)
        (PP (IN about)
         (NP (NN seventeen)))
        (PP (IN below)
         (ADVP (RB up) (RB here))))))))
   (CC and)
   (S
    (ADVP (RB then))
    (NP (PRP he))
    (VP (VBZ calls)
     (NP (PRP us))))
   (. .)))
```

Figure 2: Example of the syntactic analysis of the dialogue turn that appears in Figure 1.

3.2 Syntactic features and segmentation

As we are interested in studying the coincidence of syntactic structures with DA segments, we will select two general features for each word (see Figure 3):

- Most general syntactic category that starts with a word, **(MGSS)**, i.e., the root of the current subtree of the syntactic analysis, (e.g. in Figure 2, ”CC” is the MGSS of the first word of the second segment, ”and”).
- Most general syntactic category that ends with a word, **(MGSE)**, i.e., the root of the subtree of the syntactic analysis that ends with that word, (e.g. in Figure 2, ”S” is the MGSE of last word of the first segment, ”here”).

Using these features, we have analysed the syntactic categories of boundary words of segments. Particularly, it seems interesting to study MGSE of last word of the segment and MGSS of first word of the segment, because it indicates which syntactic structure ends before the segment boundary and which one starts after it. As there is always the beginning of a segment with the first word of the turn and the end of a segment with the last word of the turn, we are ignoring these for the analysis, because we are looking for intra-turn segments. Results of this analysis can be seen in Table 1.

4 The model

The statistical model used to DA label and segment the dialogues is extensively explained in (Martínez-Hinarejos (2008)). Basically, it is

ROOT+-+: **$LAUGH** S+**he**+NP VP+**waits**+VBZ SBAR+**until**+IN S+**it**+NP VP+**gets**+VBZ PP+**about**+IN NP+**seventeen**+PP PP+**below**+IN ADVP+**up**+RB RB+**here**+S **$SEG** CC+**and**+CC S+**then**+ADVP NP+**he**+NP VP+**calls**+VBZ NP+**us**+S .+**.**+ROOT **$SEG**

Figure 3: For each word of the example turn of Figure 1, MGSS (item before the word) and MGSE (item after the word) are obtained from the tree of Figure 2. Non-verbal labels were reincorporated.

MGSE			MGSS		
Occ	%	Cat	Occ	%	Cat
33516	37.1	,	30318	33.5	ROOT
30640	33.9	ROOT	19988	22.1	CC
7801	8.6	:	13275	14.7	NP
7134	7.9	S	10187	11.3	S
2687	3.0	NP	3508	3.9	SBAR
2319	2.6	PRN	3421	3.8	ADVP
750	0.8	VP	2034	2.2	VP
531	0.6	ADVP	1957	2.2	INTJ
478	0.5	PP	1300	1.4	UH
465	0.5	RB	972	1.1	PP
4078	4.5	Other	3481	3.8	Other

Table 1: Occurrences and percentage of the syntactic categories that correspond with the most frequent MGSE of the last segment word (except last segment) and MGSS of the first segment word (except first segment).

based on a combination of a Hidden Markov Model at lexical level and a Language Model (n-gram) at DA level. The Viterbi algorithm is used to find the most likely sequence of DA labels according to the trained models. The segmentation is obtained from the jumps between DAs of this sequence.

The previous section has shown that the MGSE and MGSS for the segments boundary words are concentrated in a small set of categories (see Table 1). Therefore, one quick and easy way to incorporate this information to the existing model is to add some restrictions during the decoding process, giving the model:

$$\widehat{U} = \arg\max_{U} \max_{r,s_1^r} \prod_{k=1}^{r} \Pr(u_k|u_{k-n-1}^{k-1}) \cdot \\ \cdot \Pr(W_{s_{k-1}+1}^{s_k}|u_k)\sigma(x_{s_k})$$

where $\widehat{U}$ is the sequence of DAs that we will get from the annotation/segmentation process. The search process produces a segmentation $s = (s_0, s_1, \ldots, s_r)$, that divides the word sequence W into the segments $W_{s_0+1}^{s_1} W_{s_1+1}^{s_2} \ldots W_{s_{r-1}+1}^{s_r}$. Each segment is assigned to a DA u_i that forms the DA sequence $U = u_1 \ldots u_r$. x_i corresponds to the syntactic features of the i word that can be MGSE, MGSS or both of them, and

$$\sigma(x_i) = \begin{cases} 1 & \text{if } x_i \in X \\ 0 & \text{otherwise} \end{cases}$$

where X can be a subset of all the possible syntactic categories that correspond to:

1. the most frequent MGSE of last segment word, if x is MGSE.
2. the most frequent MGSS of first segment word, if x is MGSS
3. the most frequent combinations of both previous sets.

It means that we will only allow a segment ending when the MGSE of a word is in this set, or a start of a segment when the MGSS of the following word is in the corresponding set or both conditions at the same time.

5 Experiments and results

Ten cross-validation experiments were performed for each model using, in each experiment a training partition composed of 1136 dialogues and a test set of 19 dialogues, as in (Stolcke et al (2000); Webb et al (2005); Martínez-Hinarejos et al (2006)). The N-grams were obtained using the SLM toolkit (Rosenfeld (1998)) with Good-Turing discounting and the HMMs were trained using the Baum-Welch algorithm. We use the following evaluation measures:

- To evaluate the labelling, we use the DA Error Rate (equivalent to Word Error Rate) and the percentage of error labelling of whole turns.
- For the segment evaluation, we only check where the segments bounds are produced (word position in the segment), making use of F-score obtained from precision and recall.

The results from using different sizes for the set X are shown for labelling performance in Tables 2 and 3, and F-score of the segmentation in Table 4.

Model/SizeX	5	10	20	All
MGSE	53.31	54.76	54.60	54.76
MGSS	53.35	**52.76**	54.92	54.76
Both	53.58	52.84	54.76	54.76

Table 2: DAER for models using MGSE, MGSS and both features. SizeX indicates the size of the set of most frequent categories accepted. Without syntactic categories (baseline) we obtain a DAER of 54.41.

Model/SizeX	5	10	20	All
MGSE	53.61	55.41	55.34	55.77
MGSS	53.61	53.32	55.63	55.77
Both	53.46	**53.10**	55.19	55.77

Table 3: Percentage of error of labelling of complete turns for all the possible models. The baseline value is 55.41.

Model / SizeX	5	10	20	All
MGSE	73.08	71.18	71.44	71.17
MGSS	73.60	73.72	71.44	71.17
Both	**74.36**	74.08	71.75	71.16

Table 4: F-score of segmentation. The baseline value is 71.17.

6 Discussion and future work

In this work, we have used lexical and syntactic features for labelling and segmenting DAs simultaneously. Syntactic features obtained automatically were deterministically applied during the statistical decoding process. There is a slight improvement using syntactic information, obtaining better results than reported in other work such as (Martínez-Hinarejos et al (2006)). The F-score of the segmentation improves 3% using the syntactic features, however values are slightly worse (2%) than results in (Stolcke and Shriberg (1996)).

As future work, we think that incorporating the syntactic information in a non-deterministic way might further improve the annotation and segmentation scores. Furthermore, it is possible to make use of additional information from the syntactic structure, rather than just the boundary information we are currently using. Finally, an evaluation over different corpora must be done to check both the performance of the proposed model and the reusability of the syntactic sets.

Acknowledgments

This work was partially funded by the Companions project (http://www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

References

Ang J., Liu Y., Shriberg E. 2005. Automatic Dialog Act Segmentation and Classification in Multiparty Meetings. Proc. ICASSP, Philadelphia, USA, pp. 1061-1064

Ji, G and Bilmes, J. 2005. Dialog act tagging using graphical models. Proc. ICASSP, Philadelphia, USA

Jurafsky, D. Shriberg, E., Biasca, D. 1997. Switchboard swbd-damsl shallow- discourse-function annotation coders manual. Tech. Rep. 97-01, University of Colorado Institute of Cognitive Science

Klein D. and Manning, C. D. 2003. Accurate Unlexicalized Parsing. Proc. ACL, Sapporo, Japan, pp. 423-430

Martínez-Hinarejos, C. D., Granell, R., Benedí, J. M. 2006. Segmented and unsegmented dialogue-act annotation with statistical dialogue models. Proc. COLING/ACL Sydney, Australia, pp. 563-570

Martínez-Hinarejos, C. D., Benedí, J. M., Granell, R. 2008. Statistical framework for a spanish spoken dialogue corpus. Speech Communication, vol. 50, number 11-12, pp. 992-1008

Rosenfeld, R. 1998. The cmu-cambridge statistical language modelling toolkit v2. Technical report, Carnegie Mellon University

Stolcke, A. and Shriberg, E. 1996. Automatic linguistic segmentation of conversational speech. Proc. of ICSLP, Philadelphia, USA

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M. 2000. Dialogue act modelling for automatic tagging and recognition of conversational speech. Computational Linguistics 26 (3), 1-34

Webb, N., Hepple, M., Wilks, Y. 2005. Dialogue act classification using intra-utterance features. Proc. of the AAAI Workshop on Spoken Language Understanding. Pittsburgh, USA