In Proceedings of ACL-99 Workshop on Discourse Tagging

Tagging of Speech Acts and Dialogue Games in Spanish Call Home

Lori Levin, Klaus Ries, Ann Thymé-Gobbel, Alon Lavie
Language Technologies Institute
Carnegie Mellon University
Internet: lsl@cs.cmu.edu

Abstract

The Clarity project is devoted to automatic detection and classification of discourse structures in casual, non-task-oriented conversation using shallow, corpus-based methods of analysis. For the Clarity project, we have tagged speech acts and dialogue games in the Call Home Spanish corpus. We have done preliminary cross-level experiments on the relationship of word and speech act n-grams to dialogue games. Our results show that the label of a game cannot be predicted from n-grams of words it contains. We get better than baseline results for predicting the label of a game from the sequence of speech acts it contains, but only when the speech acts are hand tagged, and not when they are automatically detected. Our future research will focus on finding linguistic cues that are more predictive of game labels. The automatic classification of speech acts and games is carried out in a multi-level architecture that integrates classification at multiple discourse levels instead of performing them sequentially.

Introduction

The Clarity project is devoted to automatic detection and classification of discourse structures in casual, non-task-oriented conversation using shallow, corpus-based methods of analysis. The main application of the Clarity discourse classifiers will be automatic detection of what discourse participants are doing. This includes identifying genres (e.g., lecture vs. casual conversation) [1], functional activities (e.g., planning vs. gossiping), and discourse phases (e.g., opening vs. closing of a negotiation) among other things [3].

The first stage of the Clarity project involved developing an architecture for a multi-level discourse classifier to be trained and tested on three discourse levels: speech acts, dialogue games, and functional activities. Tagging, training, and testing was carried out on the Call Home Spanish (CHS) corpus of casual, non-task oriented conversation between family members. The coding scheme for speech acts and games is described in [6, 9]. The automatic classification of speech acts and games is carried out in a multi-level architecture that integrates classification at multiple discourse levels instead of performing them sequentially. The multi-level classifier is described in [7].

The focus of this paper is on the cross-level classification of speech acts and dialogue games as it is carried out by the multi-level classifier. Results of a preliminary experiment show that the type or label of a game cannot be predicted from n-grams of the words it contains. Predicting game labels from the sequences of speech acts they contain yields slightly better results (better than baseline), but only when the speech acts are hand tagged, not when they are automatically detected. This is very preliminary research, and we plan in the future to experiment with linguistic features that are more predictive of game and functional activity labels.

In the course of discussing our tagging scheme and classification results, we will comment on the specific adaptations of our coding scheme for non-task oriented dialogue and for its intended application to functional activity detection.

The coding schemes for speech acts and dialogue games along with the adaptations for dialogue type and application are summarized in Sections 2 and 3. Section 4 describes the multi-level classification architecture and our experiments with cross-level classification of speech acts and games. Finally we will discuss the implications of our experiment for predicting higher levels of discourse structure from lower levels.

Speech Act Tags

Our speech act coding is based on the SWBD-DAMSL tag set from the Johns Hopkins LVCSR Summer Workshop 1997 [5]. Modifications were made in the tag set for four reasons. First, we made some changes to improve inter-coder agreement. After two taggers tagged CHS, we calculated intercoder agreement (79%. See [6] for a description of how this was computed) and examined confusion matrices showing the tags that they disagreed on most. We eliminated the tag **sv** for statement of opinion because taggers did not distinguish reliably between **sv** and **s**, plain statement. To account for some aspects of opinions that we felt could be tagged reliably, we have added features such as **val +/-** (positive or negative value judgement) to the category **s**.

The second reason for changes to the coding scheme was breaking up large categories. Language models for large categories are thought to be mixed bags of diverse information that are not sufficiently distinct from language models of other categories. Smaller categories should yield more distinct language models. In our first pass tagging of CHS, statements (**s**) accounted for around 40 per cent of the tags. We broke up the **s** category by adding semantic feature tags (certainty, hypotheticality, expresses speakers mental state, positive/negative value judgement, obligation, past vs non-past tense, joke/sarcasm). In the revised coding scheme, unmarked **s** accounts for 26% per cent of the tags.

Unfortunately, this subdivision of **s** did not work as we expected. When our automatic speech act classifier was trained on the subcategories of **s**, we found that the marked subcategories could be reliably distinguished from each other, but could not reliably be distinguished from unmarked **s**. We feel that this may be because the language model for unmarked **s** still characterizes too much diverse information which overlaps with the characteristics of the marked subcategories of **s**.

The third reason for revisions of the coding scheme was to support the task of functional activity identification. We hypothesized that semantic features such as hypotheticality and uncertainty, might characterize planning-type activities in contrast to gossip-type activities. The correlation between our semantic features and functional activities remains to be tested.

The fourth reason for revising the coding scheme was to account for types of utterances that were very prominent in CHS, but not so prominent in SWBD. One tag that we added for this reason was **atd** (attention directive) for cue phrases such as *oye* (listen) and *mira* (look).

Our speech act tags fall into the categories Questions, Answers, Agreement/Disagreement, Discourse Markers (including Backchannels), Forward Functions, Control Acts, Statements, and Other. The greatest changes with respect to SWBD-DAMSL involve an expansion of control acts, the addition of coarse emotional tags to backchannels, a collapse of statements and opinions, and the addition of semantic features to statements. The complete speech act tag set is listed in [6, 9].

The entire CHS corpus was re-tagged by one human tagger using this revised tag set. Intracoder agreement of over 90% was tested by having the tagger re-tag three dialogues several weeks after first tagging them. (Again, see [6] for how agreement was calculated.)

Dialogue Game Tags for Non-Task-Oriented Dialogue

We use the following characterization of dialogue games given by Carletta at al. [2]: "a set of utterances starting with an initiation and encompassing all utterances up until the purpose of the game has been either fulfilled (e.g. the requested information has been transferred) or abandoned." However, our inventory of games differs from those used for the Map Task because we are dealing with non-task oriented dialogue. Our system contains eight main types of games plus eight modifiers. The game types are **seeking information**, **giving information**, **giving directive**, **action commit**, **giving opinion**, **expressive**, **seeking confirmation**, and **communication filler**. The eight modifiers of game tags are **sup** (support previous statement), **elab** (elaborated answer to question), **spec** (speculation), **wk** (weak directive or commit), **st** (strong directive or commit), **n** (nested game), **inc** (incomplete game), **aband** (abandoned game). Taggers label turns within a game as **Initiative**, **Response**, and **Feedback**. Figure 1 shows a fragment of a CHS dialogue tagged with speech acts and games.

Games may overlap, either as nested games or as interleaved games. Game boundaries are determined by changes in who has the initiative and changes in speaker intention, for example changing

```
#Game:Info
<I> s A: Llevan una lavadora,
         they bring a washing machine
<I> s A: llevan una no se' que',
         they bring I don't know what
<I> s A: llevan una cosa de cada cosa
         they bring one of each
#Game:Quest^Aband
<I> qw B: pero como,
          but how
#Game:Quest
<I> qy B: pero pagan impuestos,
          but are they taxed
<I> s^cert-
    B: pero se supone que el menaje no paga
    but household items are not supposed
    to be taxed
<R> ny A: si'
          yes
#Game:Info^Elab
<I> s^e A: no si' paga impuestos,
           no yes it is taxed
<I> s^cert+
    A: paga el quince por ciento, si' sen~or
       it's taxed fifteen per cent, yes sir
<R> b  B: ah si'
          oh yes
<R> s^m B: paga quince por ciento de impuesto
           it's taxed fifteen percent
```

Figure 1: A Fragement of a Tagged Dialogue

from informing to questioning.

The Multi-Level Classifier Architecture

The detection of speech acts from words has received a substantial amount of attention in the recent past. Most of the recent literature (e.g. see [8, 4] and their references) describes classifiers that are based on n-gram modeling. This technology has proven to give reasonable results in a variety of domains, although some recent papers indicate shortcomings of this approach [10, 7].

Detection from the word level: It is not obvious from the outset how n-gram-based classifiers should be applied to the detection of discourse games. One could try to do segmentation and labeling of discourse games directly from the word level just as in the speech act task. This approach can be characterized as:

$$\mathbf{G}^* = \text{argmax}_{\mathbf{G}} p(\mathbf{G}|\mathbf{W}) = \text{argmax}_{\mathbf{G}} p(\mathbf{W}|\mathbf{G}) p(\mathbf{G})$$

Assuming that we can model $p(\mathbf{G})$ reasonably well by an n-gram model, the remaining task is to find a good estimate for $p(\mathbf{W}|\mathbf{G})$. Assuming that the words in a game segment (given the game type) do not depend on other games or words this can be simplified to the more local rule:

$$p(\mathbf{W}|\mathbf{G}) = \prod_{i=1,\ldots,n} p(\mathbf{W}_i|\mathbf{G}_i)$$

Furthermore – to make the model tractable – we would like to model these probabilities with n-gram models. However, it is highly questionable that $p(\mathbf{W}_i|\mathbf{G}_i)$ can adequately be modeled using standard n-gram models, because:

- it appears that a much smaller portion of n-grams have discriminative distributions for dialogue games in comparison with the speech act case
- the interaction between the speakers is lost using word level bigram or trigram statistics

Detection from the speech act level: For these reasons, we have been pursuing a second possible modeling approach, which estimates $p(\mathbf{W}_i|\mathbf{G}_i)$ with a probability model that uses speech-acts as an intermediate discourse level:

$$p(\mathbf{W}_i|\mathbf{G}_i) = \sum_S p(\mathbf{W}_i|\mathbf{S}_i, \mathbf{G}_i) p(\mathbf{S}_i|\mathbf{G}_i)$$

Using similar independence assumptions as above we can rewrite this as

$$p(\mathbf{W}_i|\mathbf{G}_i) = \sum_S p(\mathbf{S}_i|\mathbf{G}_i) \prod_{j=1,\ldots,m_i} p(\mathbf{W}_{i,j}|\mathbf{S}_{i,j}, \mathbf{G}_i)$$

In this setup potential speech act constraints of dialogue games are enforced explicitly. The distributions of $p(\mathbf{W}_{i,j}|\mathbf{S}_{i,j}, \mathbf{G}_i)$ are also likely to be sharper since they are shorter and do not mix unrelated grammatical information. To make this model more tractable we can use a Viterbi approximation (which finds and uses the most likely speech act sequence rather than summing over all possible such sequences). Another simplification is to assume that the word probabilities are not dependent on the dialogue game they occur in, as long as the speech act is known:

$$p(\mathbf{W}_{i,j}|\mathbf{S}_{i,j}, \mathbf{G}_i) = p(\mathbf{W}_{i,j}|\mathbf{S}_{i,j})$$

This assumption can in fact be partially removed for frequent speech acts (for example, one could have separate statement types for different dialogue games). The overall estimate for $p(\mathbf{W}_i|\mathbf{G}_i)$ would therefore be:

$$argmax_S \prod_{i=1,\ldots,n} p(\mathbf{S}_i|\mathbf{G}_i) \prod_{j=1,\ldots,m_i} p(\mathbf{W}_{i,j}|\mathbf{S}_{i,j})$$

Incremental Lattice Construction: Both models (almost) fulfill the form of a recursive HMM, and a dynamic programming solution is known for finding the Viterbi speech-act path. However the algorithm is complex, especially if higher order models are used. Additionally, this model ignores the problem of modeling the overlap between speakers, which, when attacked, further complicates the algorithm. It would also be nice to be able to use more complex models for dialogue game detection, especially models that would allow us to transcend the usual n-gram constraints and that could be trained discriminatively [1]. Given these complications, we had previously proposed an incremental lattice construction approach [7] (see Fig. 2), that does the following:

1. generate a lattice from the words
2. generate a lattice of all likely segments from the lattice of the previous level
3. label all segments in the new lattice with their tag and likelihood
4. goto step 2 (next discourse level) or continue to step 5
5. search the most likely path in the complete lattice given a language model and the likelihoods in the lattice

The input to the discourse classifier can be a lattice of words, e.g. produced by a speech recognizer. A lattice of segments is produced by an A^* search over the input lattice that can hypothesize segment boundaries and inserts every segment hypothesis into the segment lattice. The speech act lattice is generated by replacing each segment by the set of all possible speech act labels and assigning a likelihood for each segment given the speech act. This lattice can then be segmented again and annotated with dialogue games. A Viterbi search can then find the best combined speech act and game sequence.

Experiments We have conducted experiments for dialogue game detection where we assumed that the boundaries of the games were known [2].

We used the seven primary game tags **info**, **quest**, **seek_conf**, **opinion**, **direct**, **express** and **UNDEFINED**. Since the primary tag **info** was frequently combined with the orthogonal tag **elab** we split the **info** category along that secondary tag. We used 22 dialogues for training and testing, applying a 7-fold cross-validation. No intercoder agreement tests have been conducted so far (Table 1).

Game Label Detection	
Input	Accuracy
chance	40.7%
manual speech acts	55.0%
words	40.7%
speech acts detected from words	46.5%

Table 1: Dialogue game detection results for a given dialogue game segmentation

Discussion and Future Directions

The results of our dialogue game detection experiment (Table 1) indicate a number of interesting conclusions. First, game tags cannot be predicted from n-grams of words — at least not taking all words into account equally. This does not eliminate the possibility that there may be discriminative n-grams of words hidden in various positions in the game, and this is a research direction that we would like to explore. Our second observation is that speech acts help a little in predicting game labels, although current speech act classification technology does not yield good enough results for this purpose. The reason that speech acts help as an intermediate level between words and games may be because the speech acts already encode some of the discriminating words that are necessary for game detection. Finally, we are left with the observation that we have not yet identified the linguistic cues that are relevant to game classification. Our future plans include experimenting with a variety of linguistic cues of the sort discussed

[1] The models mentioned above can easily be trained with the maximum likelihood criterion. However – since we have many irrelevant n-grams – it might be advantageous to train these models just to attain a good discrimination performance, see [7, 10].

[2] The main problem is that one has to deal with overlap in a slightly different way than in the speech act case.

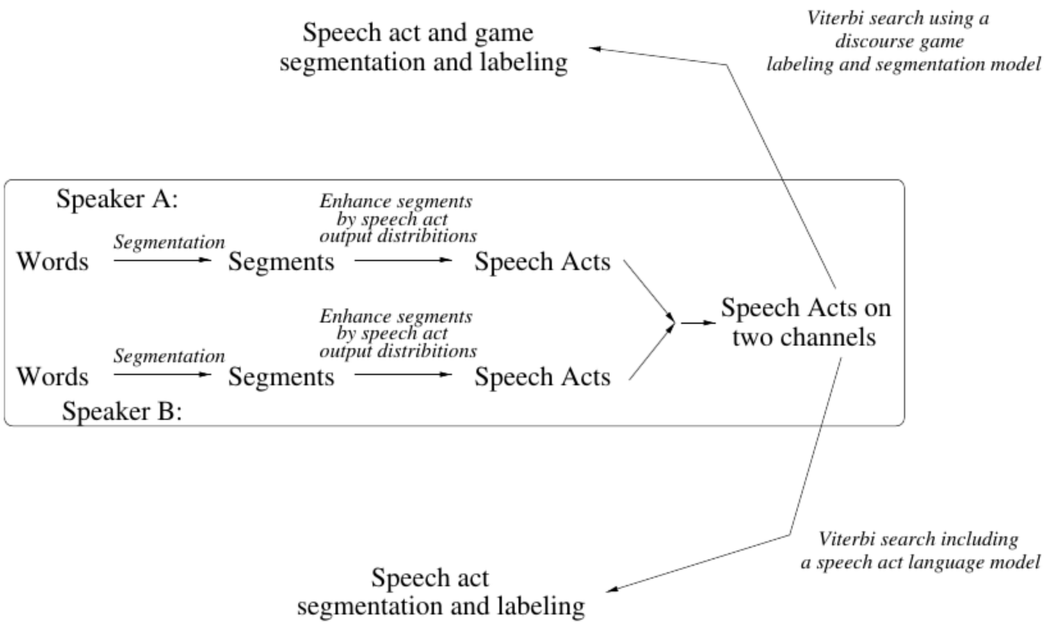
Speech act and game segmentation and labeling
Viterbi search using a discourse game labeling and segmentation model
Speaker A:
Words
Segmentation
Segments
Enhance segments by speech act output distribitions
Speech Acts
Speech Acts on two channels
Words
Segmentation
Segments
Enhance segments by speech act output distribitions
Speech Acts
Speaker B:
Viterbi search including a speech act language model
Speech act segmentation and labeling

Figure 2: Incremental lattice construction (reproduced from [7])

by [3, 1] in the hope of finding some predictive features that are not captured by our current speech act tag set.

Another line of research is to question the relevance of the discourse units we are using — dialogue games in particular — to the tasks of identifying functional activities, genres, or phases of a discourse. On the one hand, dialogue games provide potentially important information about which utterances are initiatives and which are responses, thus helping to identify which speakers are dominant at different points in the conversation. On the other hand, dialogue games may not identify initiative in the most useful way. If the first speaker is providing a recount of an event the second speaker may often ask a question with the sole purpose of displaying his/her attention to the story being told. Our current tagging scheme would identify such a question as an initiating utterance of a question/answer game that is initiated by the second speaker with the first speaker in a responding role. Clearly this does not reflect the actual discourse situation in this case: The *first* speaker is in charge and the second speaker merely displays attention.

For reasons such as this, we have begun to investigate different types of discourse units, such as genre units which consist of utterances having genre-specific functions. A genre unit might, for example, be a piece of a conversation that contains gossip. Utterances in this gossip unit might have gossip-specific functions such as introducing a topic of gossip, accepting or rejecting the topic, giving the main value judgement, etc. Another genre unit might be a story and one of the genre specific functions of a question might be to encourage the story teller.

Our multi-level classification architecture will help us explore alternative tagging schemes for multiple levels of discourse structure and the interaction between them.

Other Applications of the Multi-level Classifier

We are currently also using the multi-level classifier architecture described above for the task of detecting dialogue acts in the context of a spoken language translation system for the travel domain [11]. In this system, dialogue acts, parameterized with arguments and their values, serve as a shallow interlingua representation for translating spoken utterances into other target languages. To detect dialogue acts, we use a two-level model of discourse structure where the lowest level of structure consists of semantic argument categories (such as time and date expressions, location expressions, and other descriptions of travel domain entities). These semantic argument categories are detected from the utterance using word n-gram models. Speech acts serve as the second level of discourse structure in this system, and are modeled via n-grams of semantic argument categories. The combination of speech act and semantic argument categories are then mapped into a complete dialogue act. Statistical dialogue act detection is intended to serve as robust alternative to a more traditional grammar based analysis component. This indicates that the multi-level classifier architecture describes here is useful for a variety of applications involving classification of complex language and discourse structures.

References

[1] Douglas Biber, Susan Conrad, and Randi Reppen. *Corpus Lingusitics: Investigating Language Structure and Use.* Cambridge University Press, 1998.

[2] Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31, March 1997.

[3] Suzanne Eggins and Diana Slade. *Analysing Casual Conversation.* Cassell, 1997.

[4] Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner. Clarity: Automatic discourse and dialogue analysis for a speech and natural language processing system. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, March 1998.

[5] Dan Jurafsky, Liz Shriberg, and D. Biasca. Switchboard-damsl labeling project coder's manual. Technical report, Institute of Cognitive Science, University of Colorado, Boulder, USA, 1997.

[6] Lori Levin, Ann Thymé-Gobbel, Klaus Ries, Alon Lavie, and Monika Woszczyna. A discourse coding scheme for conversational spanish. In *ICSLP*, 1998.

[7] Klaus Ries. Hmm and neural network based speech act detection. In *ICASSP*, 1999.

[8] Andreas Stolcke, Elizabeth Shriberg, Rebecca Bates, Noah Coccaro, Daniel Jurafsky, Rachel Martin, Marie Meteer, Klaus Ries, Paul Taylor, and Carol Van Ess-Dykema. Dialog act modeling for conversational speech. In *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, March 1998.

[9] Ann Thymé-Gobbel and Lori Levin. *Dialogue Activity Tagging Manual for Spanish Conversational Speech.* `http://www.cnbc.cmu.edu/~gobbel/clarity/manualintro.html`, 1998.

[10] V. Warnke, S. Harbeck, H. Niemann, and E. Nöth. Discriminative estimation of interpolation parameters for language model classifiers. In *ICASSP*, 1999.

[11] M. Woszczyna, M. Broadhead, D. Gates, M. Gavaldà, A. Lavie, L. Levin, and A. Waibel. A modular approach to spoken language translation for large domains. In *AMTA-98*, 1998.