Applying Machine Learning to Discourse Processing,
AAAI'98 Spring Symposium Series, March 23-25, 1998, Stanford University, California

CLARITY: INFERRING DISCOURSE STRUCTURE FROM SPEECH

Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, Klaus Zechner
{finkem,mirella,alavie,lsl,laura,tpolzin,ries,ahw,zechner}@cs.cmu.edu

Interactive Systems Labs
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA

ABSTRACT

The goal of the CLARITY project is to explore the use of discourse structure in the understanding of conversational speech. Within project CLARITY we aim to develop automatic classifiers for three levels of discourse structure in Spanish telephone conversations: speech acts, dialogue games, and discourse segments. This paper presents our first results and research plans in three areas: definition of discourse structure units and manual annotation of CALLHOME SPANISH, speech recognition, and automated segmentation and labeling of speech acts.

1. INTRODUCTION

This paper describes preliminary work on the CLARITY project at Carnegie Mellon University (CMU). The goal of the CLARITY project is to explore the use of discourse structure in the understanding of conversational speech. Our role in the CLARITY project is to develop automatic segmenters and labelers for three levels of discourse structure in Spanish telephone conversations. Automatic segmenters and labelers for speech acts, dialogue games, and discourse segments will be developed and tested on transcribed data as well as speech recognizer output. We will be using the CALLHOME SPANISH corpus.

The major research issues we address are: (1) the definition of discourse structure units that are applicable to non-task-oriented dialogue, (2) the design of a unified architecture for discourse structure classifiers at different levels (speech acts, dialogue games, and discourse segments), (3) the role of prosody in the automatic detection of discourse structure, and (4) the integration of discourse structure classification and state-of-the-art Large Vocabulary Continuous Speech Recognition (LVCSR).

The CLARITY project and the work reported in this paper are funded in part by a grant from the US Department of Defense.

This paper presents our first results and research plans in three areas: definition of discourse structure units and manual annotation of CALLHOME SPANISH, speech recognition, and automated segmentation and labeling of speech acts.

2. MANUAL DISCOURSE ANNOTATION

An initial task of the CLARITY project is the manual annotation of discourse structure in CALLHOME SPANISH. The manual tags will serve as training data for the automatic discourse classifiers. We will be primarily interested in three levels of discourse structure: speech acts, dialogue games ([CII^{+}97]), and discourse segments ([Hea97, NGAH95, PL97]). The manual annotation consists of defining these levels of discourse structure in a way that (1) is applicable to CALLHOME SPANISH and (2) can be tagged with a high degree of intercoder agreement.

Two main approaches have been proposed for speech act annotation. The DAMSL (Dialogue Act Markup in Several Layers) annotation scheme ([CA97]) was primarily developed for two-agent task oriented dialogues, in which participants collaborate to solve some problem. The main feature of this scheme is that it attempts to capture the utterance's purpose by allowing multiple labels in multiple layers for a single utterance. Thus, an utterance might simultaneously perform actions such as responding to a question, confirming understanding and informing. Our experiments with the DAMSL annotation scheme have shown that it is difficult to apply to non-task-oriented dialogues.

The Johns Hopkins ([JSB97]) speech act inventory was developed for non-task-oriented dialogues and used for the annotation of the Switchboard English corpus. The Johns Hopkins tagset was designed so as to be compatible with the DAMSL annotation scheme. It does not allow, however, for hierarchical annotation of discourse structure.

We have so far concentrated on extending and modifying the Johns Hopkins tagset in order to capture regularities of the CALLHOME SPANISH corpus. Although preliminary, the application of the Johns Hopkins tagset on the CALLHOME data revealed the following: (1) the tagset needs to be extended so as to account for the distinction between direct and indirect speech (e.g., *entonces él me dice que en, que en diciembre las tarifas son más bajas* 'he told me that in December the prices are much lower'), expressions of surprise (*ajá* 'ah', *uy qué caray* 'good grief', *qué lástima* 'what a pity'), and attention directives (*oye* 'listen', *mira* 'look'); (2) some of the Johns Hopkins tags simply do not occur in the current fragment of the data we have been experimenting with (for example, re-formulations of the interlocutor's utterance and self-talk); and (3) with about sixty separate speech act tags, the tagset is hard to learn.

At a higher level of discourse structure we plan to identify and manually annotate dialogue games and discourse segments. Dialogue games ([Car83, LM77, Pow79]) center around an initiating utterance and a responding utterance. Examples are questions followed by answers, statements followed by acceptance or denial, etc. The sequence of speech acts within a dialogue game defines a *discourse state*. For example, a negotiation state consists of a suggestion followed by a rejection or acceptance and an information seeking state consists of a question followed by an answer. Sequences of discourse states in turn will indicate *dialogue types*. For instance, a dialogue consisting of sequences of information seeking and negotiation games could be a negotiation dialogue.

Several prominent theories of discourse assume a hierarchical segmentation model. Among these are the attentional/intentional structure of ([GS86]) and the Rhetorical Structure Theory of ([MT88]). The building blocks of these theories are phrasal and clausal units and have been mainly applied to short texts, typically one to three paragraphs in length. Our initial experiments show that it is not an easy task to annotate hierarchical segments, the reason being that most CALLHOME dialogues consist of many linear segments. Concerning the manual tagging of topic shifts we intend to address the following research issues: (1) defining the notion of dialogue sub-topics for spoken language where a dialogue may drift between sub-topics without clear boundaries and without an all-encompassing main topic and (2) ensuring intercoder reliability on topic segmentation ([PL97]).

We are also interested in sociolinguistic and situational factors in discourse. This would include both detection of social features that affect discourse and in describing the social situation that people are in by analyzing the discourse. ([Cla96]). For example, in interactions between family members, which make up most of the CALLHOME conversations, the social work of maintaining a relationship and face guide the conversation. We also intend to characterize conversations in terms of their emotional status as well as the level of confrontation in the conversation.

3. AUTOMATIC DISCOURSE COMPONENTS

The target of this investigation is to build a fully automatic model of discourse structure. Our automatic discourse classifier will take input *features* such as the lexical items in an utterance, the intonation contour, etc. Techniques for extracting the relevant features are described in Sect. 3.1 (for identifying words) and Sect. 3.2 (for prosodic feature extraction). We also plan to use cue phrases and other derived features ([WGR97, Hea97]).

Since the layers of discourse structure are interdependent each other, it will be necessary to integrate the discourse classifiers for speech acts, dialogue games, and discourse segments. Therefore we need a search engine that considers all possible labelings and segmentations in parallel. For this major undertaking we are following a step-by-step approach:

1. build a speech act segmenter and a speech act labeler that are separate from each other
2. build an integrated segmenter and labeler for the speech act level with a simple search procedure
3. develop classifiers for higher levels (games and segments) assuming perfect classification on the lower levels
4. build a search engine that takes all three levels of discourse structure into account

3.1. Speech Recognition

3.1.1. The task

CALLHOME SPANISH is a database provided by the Linguistic Data Consortium (LDC) consisting of 120 unscripted dialogues between native speakers of American Spanish. The telephone calls originated in the United States, and speakers were permitted to place calls to foreign locations to speak with their families or close friends for up to 30 minutes with the understanding that their speech would be recorded and used for research. Five to ten minutes of each dialogue were transcribed. Topics and word choice were completely unrestricted.

Callers were speaking to callees with whom they were very familiar, leading to an entirely different speaking style from that seen in other databases (see Table 1). The speakers jump from topic to topic, skipping the kinds of formalities expected in discourse between strangers. The channel at the foreign site was often shared between several people, with people talking at once and mothers calling for sisters to put the baby on the line. Speech is sometimes directed at a third person at the foreign site, but is still passed through the channel. Pronunciation is very familiar, with little care paid to articulation. Dialect differences (Carribean, Mexican, Río Plata among others) are far more pronounced than in English databases. There is also a high occurrence of non-Spanish words (English, Indian dialects). Although our own experience has not shown recognition of Spanish to be significantly more difficult than that of English, the data we have used in the past has been more homogeneous than that in CALLHOME SPANISH. State-of-the-art performance for CALLHOME SPANISH has been reported with 57% word error ([ch-97]).

3.1.2. CALLHOME SPANISH *put in context*

CALLHOME SPANISH is one of the most difficult tasks currently being undertaken by the speech recognition community. The main sources of difficulty inherent to the task are the style of speech (spontaneous, unrestricted), variability of dialect, and quality of the channel (telephone band; varying quality of foreign line). Additionally, there are challenges such as sparseness of data that also must be addressed.

CALLHOME SPANISH may be best characterized in the context of other tasks (Table 1).

task	style	restrictions	speakers	WA
WSJ	read	exact text	NA	95%
SST	spont.	one topic	colleagues	80%
SWB	spont.	one of 70 topics	strangers	65%
CH	spont.	unrestricted	family	45%

Table 1: Comparison of English LVCSR tasks (abbreviations are WSJ: Wall Street Journal; SST: Spontaneous Scheduling Task; SWB: Switchboard; CH: CALLHOME; WA: Word Accuracy)

CALLHOME is *unrestricted, spontaneous* speech, which is characterized by:

- high human noise content
- unrestricted vocabulary
- frequent ungrammaticalities
- regional variations
- careless pronunciation

3.1.3. Speech Engine

The system we are currently developing is a gender-independent, continuous mixture density, tied state cross-word context-dependent hidden Markov model (HMM) system based on the Janus Recognition Toolkit ([FGH+97]). Training of the new system was initiated by bootstrapping from a Serbo-Croatian system; Serbo-Croatian was chosen because it contains most of the phonemes of Spanish. We used 25 speech phones, with four noise phones and one silence phone. HMMs were 3-state; polyphonic clustering was done with a $\pm$ 2-phone context.

The CALLHOME SPANISH training data provided by LDC was augmented with recordings from Ricardo, a database of telephone monologues. Words from Ricardo were added to the LDC dictionary.

The recognizer has not yet been fully tested, but preliminary tests on the first version of the system indicate that we will soon be in the range of performance reported by other sites.

3.2. Prosodic database

Prosody plays an extensive role in the automatic identification and classification of the various levels of discourse structure. The prosodic features we consider are pitch, intensity, and speaking rate. We employ several normalization techniques to arrive at more robust features. We also use regression techniques to summarize the dynamic behavior of pitch or intensity succinctly.

These prosodic features represent an additional input source for the classifiers. Note that decisions of these classifiers are now based on multiple input modes, with inputs such as prosodic features and language models.

We use prosodic information to facilitate the segmentation of an utterance into basic speech-act-level units. Moreover, we use prosodic information as one source of information when we determine speech acts. For these purposes we generate a prosodic database which allows the classifiers fast and flexible access to prosodic information.

3.3. Automatic discourse segmentation and classification

Most work on automatic classification of speech acts has used a "Markov model" approach [JBC+97a, KKN+97, ARM97]. The work on call-type classification in [Gor95, WGR97] could be called a "direct classification" approach. Additionally we have introduced a "direct neural classification" approach [GZA97, BW96] on related classification problems. We argue here that we need a "direct hybrid classification" approach.

We will first address the problem of speech act classification in isolation, showing how we can detect the speech act labeling U given the prosodic feature vector for each utterance F and the sequence of words within that utterance W. This problem will serve as a prototypical case for different modeling approaches. If we want to refer to an individual utterance i we will write U_i, F_i and W_i. We use the *Maximum A Posteriori* (MAP) criterion, which leads to a minimum probability of error.

$$U^* = \mathrm{argmax}_U p_M(U|F, W)$$

In training, we have to find the model M^* ([DH73]) as

$$M = \mathrm{argmax}_{M'} p_{M'}(U|F, W)$$

3.3.1. *Markov Model approach*

In ([JBC+97a, KKN+97, ARM97]) a basic Markov model approach is used. The essential idea is to use Bayes theorem and then drop a "constant" term $p_M(F, W)$.

$$p_M(U|F, W) \propto p_M(F, W|U) \cdot p_M(U)$$

A problem we notice with this assumption is that $p_M(F, W)$ is not necessarily a constant during training for all classes of models.

The next assumption is that the prosodic features do not depend on the words if we know the utterance type $p_M(F|W, U) \approx p_M(F|U)$; we do not consider this assumption to be harmful in the context of the analysis we plan to do.

To result in a feasible model we have to make two further assumptions, namely $p(F|U) \approx \prod_i p(F_i|U_i)$ and $p(W|U) \approx \prod_i p(W_i|U_i)$. These assumptions are significantly more harmful since

- wrong decisions in the tagset can affect the performance of the classifier more than an approach not based on these two assumptions, since speech acts might be disambiguated by context. This might be especially true for statements ([JBC+97b]).

- we would assume that the quality of a prosodic marking depends on the "predictability" of the speech act from the discourse context

The overall model, therefore, becomes

$$p_M(U|F, W) \propto p_M(U) \prod_i p(F_i|U_i) \cdot p(W_i|U_i)$$

where $p_M(U)$, the discourse grammar, and $p(W_i|U_i)$, the speech-act-dependent language model, are typically estimated as ngram models. Even a unigram model for U, which does not even need a search, gives good results.

3.3.2. *Direct Classification approach*

The approach as presented by ([Gor95, WGR97]) does not take prosodic features into account and it only looks at isolated utterances. This roughly corresponds to using a unigram model for $p_M(U)$ which gave good results for the Markov model approach. It is therefore sufficient to estimate

$$p(U_i|W_i).$$

Under the assumption that the words in the utterance are conditionally independent, ([Gor95, WGR97]) show that a simple formula can be used to estimate $p(U_i|W_i)$. In contrast to a similar Markov model, this approach does not use higher order ngrams. ([WGR97]) further developed their approach to automatic inference of salient grammar fragments and introduced higher order ngrams in a robust way. However, the approach is fairly limited and does not allow other knowledge sources to be integrated.

3.3.3. *Direct Hybrid Classification approach and feature dependency*

This classification approach both makes use of the models that we have seen so far and extends them and builds on previous work.

neural segmentation algorithms: The neural segmentation procedure ([GZA97]) is very promising and competitive with the corresponding (hidden) Markov approach.

classification of utterances from prosodic features: In preliminary work on the SWITCHBOARD database [1] we used exponential models and neural networks to classify utterances from prosodic features and got a significant improvement in detection accuracy over the baseline CART model.

neural parsing of natural language: Two neural parsers with a *chunk and label* approach have been developed in our group with success ([Buø96, Jai91]).

Similar to our preliminary work on the prosodic database, we use neural networks with softmax ([Bri90, Jor95, Bis95]) as the activation function of the output layer. For simplicity, we will only present exponential models [2], which are essentially a version of these networks without a hidden

[1] This work was carried out on the database that was built for the LVCSR summer workshop at the CLSP at Johns Hopkins University. We are indebted to everyone in the workshop and the sponsors.

[2] Exponential models have been studied in language modeling for speech recognition and are often referred to as maximum entropy models. However, they are trained according to a different optimality criterion, the maximum entropy criterion. In our preliminary experiment we found that training with the MAP criterion was significantly more appropriate for the prosodic features.

layer. A so-called prior distribution $\hat{y}(z)$ can be used. The output of the exponential model is a probability distribution and can be written as

$$y(x, z) = exp(A \cdot x) \cdot \hat{y}(z)/Z(x, z)$$

where $Z(x, z)$ is chosen such that the output vector $y(x, z)$ sums to unity. If we choose the prior $\hat{y}(z)$ to be uniform and the input vector to be a count vector of ngrams in the utterance, this model is essentially a Markov model with a unigram discourse model ($p_M(U)$).

Using the Markov model result as the prior distribution and the trigram information as part of the input features we can construct a classifier that can benefit maximally from the good estimations of the Markov models and integrate other knowledge sources without making implicit independence assumptions.

The most prominant prosodic information source that might be already included in the Markov model is the length of the utterance (Sec. 4.3). One of our results is that the distribution is closely modeled by the Markov model and that we may be able to give a parametric distribution for this distribution. We could therefore estimate

$$\begin{aligned} &p(\text{sentence}|\text{speech act}, \text{length}) \\ &= \frac{p(\text{sentence}|\text{speech act})}{p(\text{length}|\text{speech act})} \end{aligned}$$

and use a parametric estimate for $p(\text{length}|\text{speechact})$. This approach might result in better models than the straightforward Markov model approach and if successful could eliminate the need for the full hybrid approach outlined above.

3.3.4. The segmentation problem and an integrated segmentation and labeling approach

The segmentation problem has not been explicitly addressed here so far. It can be formulated in two different ways, either as a hidden event between two words in a hidden Markov model or as a classification problem. In the first case a search must be involved, in the second case a search may be used but is not required. As ([GZA97]) shows for for speech acts, the segmentation problem itself seems to be fairly tractable even without combining it with the speech act labeling. However, if one looks at the way the segmentation and the classification model work, it becomes obvious that both of these can be integrated into one model instead of integrating them at the likelihood level. Since many word-level features for speech act classification might also be position dependent, this seems to be a simple solution and might remove false independence assumptions between the speech act segmentation and labeling models. This integration is also much cleaner than ([JBC$^+$97a, KKN$^+$97]) in conjunction with prosodic features.

4. RESULTS

4.1. Speech act segmentation

For the purpose of speech act segmentation, a time delay neural network of the standard backpropagation type was used, as is described in ([GZA97]). For training and testing we had 10 manually segmented dialogues available, containing 2221 turns. 1635 of these were used for training, 586 for testing. In total, 2983 boundaries were marked in the data, of which only 836 occurred *within* a turn, with the remaining 2147 occurring at the end of the turns.

For training, the following features were used: indication of presence of a trigger word[3] occurring in a window of ± 3 words around a (potential) speech act boundary, and the part of speech (POS) tags assigned to these words in the same window. Without an automatic POS tagger being available initially, we used a distributed encoding for the latter, in which all potential POS tags of a single word were considered to be present.

Training was performed varying the numbers of hidden units (2/4/8/10/12), as well as the number of trigger words and POS tags being used (0/10/30/50 each).[4]

The results indicate that while a high F-score[5] of $F = .84$ is achievable when measuring against *all* speech act boundaries, the performance is much worse when looking only at those boundaries that occur *within* turns: the best net yielded an F-score of .3 for those, using 4 hidden units, 50 trigger words and no POS information. We attribute the latter to the fact that the boundary-token ratio for these data is very small (less than 5 percent), meaning that the positive evidence that the neural net has to train on is very sparse.

We compared the neural segmenter to a simple Markov model approach (Fig. 1) and the experiments confirm that a POS model is important for the Markov based approach as well. The algorithm used just one language model and no search was performed; we simply assumed that there is not a speech act boundary in a window one to the left and one to the right of the boundary to be classified. Not included in the figure is an experiment on the same data set using a search algorithm without this assumption; the results are almost identical. As one can see from Fig. 1, the neural and Markov-based segmentation algorithms are both good algorithms, and depending on position on the precision/recall curve one consistently outperforms the other.

[3] A trigger word is a word occurring frequently around a speech act boundary.

[4] We always used the n most frequent trigger words or POS tags in the data.

[5] $F = \frac{2PR}{P+R}$, where P=precision and R=recall.

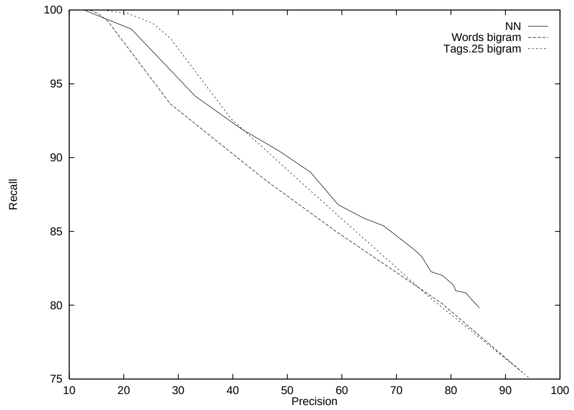

Figure 1: Performance of segmentation algorithms

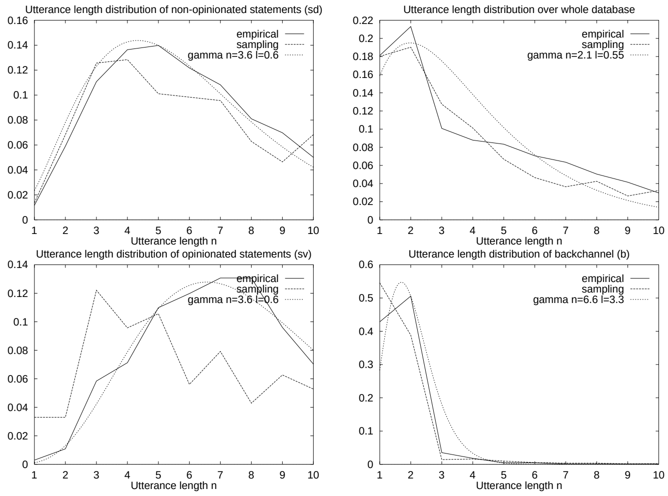

Figure 2: Length distribution of speechacts

4.2. Speech act classification

For the purpose of speech act classification, we designed a speech act classification problem similar to ([JBC+97b]). We built a tagger without a search, and used *jackknifing* to estimate accuracy with a small set of dialogues. The transcripts we used were produced by humans [6] and we compare our results to SWITCHBOARD (SWBD) ([JBC+97b]). In both cases the segmentation into utterances has been produced by humans (Table 2).

Discourse grammar	Word Model	SWBD	CALLHOME SPANISH
Chance		35%	26%
none/0-gram	3-gram	54%	41%
1-gram	1-gram		47%
1-gram	2-gram		47%
1-gram	3-gram	69%	48%
1-gram	4-gram		48%
3-gram	3-gram	71%	

Table 2: Speech act classifcation of handlabeled segments

4.3. Feature dependence

In [JBC+97b] we saw that the most salient prosodic feature for speech act classification given the segmentation is the length of the segment and that all prosodic features are highly corrolated. However [JBC+97b] assumed that the prosodic features are conditionally independent from the information derived from the LVCSR system. Fig. 2 shows that the utterance length (empirical) is fairly well modeled using the language model [7] (sampling) and that the distribution is close to the family of gamma distributions. The only exceptions are the distributions of statements which, however, comprise most of the database. This result indicates that only a feature intigration that takes the conditionally dependences into account (like the direct hybrid classification approach, Sec. 3.3.3) can fully exploit prosodic features.

4.4. Integrated classification and segmentation

We have integrated Markov model based segmentation and classification in a search procedure and applied this model to hand-transcribed data. In this case, contrary to Sec. 4.2, the segmentation has to be found by the model as well. An A^* search was used to search over the space of possible segmentations and speech act assignments and can be extend easily to support a large variety of other models such as the direct hybrid classification approach. Currently we only classify speech acts using the context one channel at a time. This results in a weaker but reasonable language model for speech act sequences for our initial investigations.

The algorithm is benchmarked using precision/recall figures for detecting the correct segment boundaries indiscriminatively of the label (segmentation) and for detecting segment boundaries with the correct label (exact match) (Fig. 3). The recall figure for the exact match is the percentage of actual segment boundaries being detected as segment boundaries and the precision is the percentage of the hypothesized segment boundaries being correctly classified. We also measured the precision (in percent correct) of the

[6] The transcriptions of CALLHOME are in our experience much less clean than those for SWBD.

[7] The language models used in this experiment are trigram models. For a unigram model theory predicts a negative expontential instead of the better fitting gamma distribution.

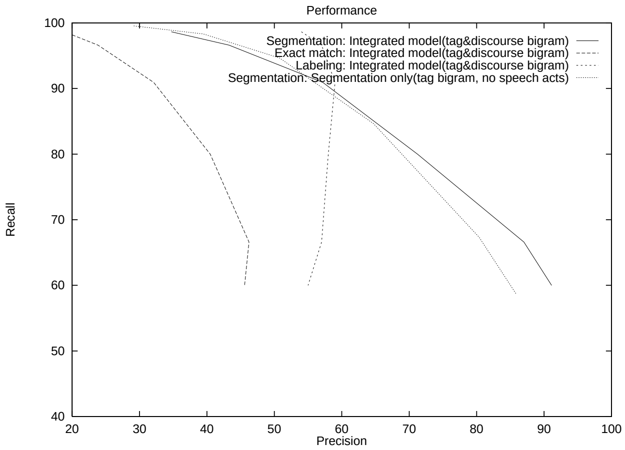

Figure 3: Performance comparision of segmentation, exact match and labeling

speech act assignment for each single word depending on the recall of the speech act boundary detection (labeling) . We varied the rate of detected speech acts using a heuristic parameter parameter to generate a precision/recall curves (Fig. 3).

Using higher order discourse models did not give us improvements, the unigram discourse model is just slightly worse in performace than the bigram discourse model. The segmentation using a segmenter without even a speech act model, similar to Fig. 1, was slightly worse than the segmenter using speech acts. This situation could change as soon as we take information form the other channel into account. Higher order speech act models were also tested but preliminary results indicate that even a unigram model delivers a good performance and we have not yet obtained higher performance with trigram speech act models.

5. SUMMARY

In this paper we presented our preliminary work and research plans on the definition of discourse structure for spoken dialogue, and the development of automatic discourse structure classifiers. In this first stage of the project, we are focusing on establishing appropriate discourse structure definitions and developing a general architecture for the segmentation and labeling tasks. Our preliminary experiments on segmentation of dialogues into speech-act-level units and on automatic labeling of these units indicate that the approach we are pursuing is both feasible and promising.

6. REFERENCES

[ARM97] Jan Alexandersson, Norbert Reithinger, and Elisabeth Maier. Insights into the dialogue processign of verbmobil. In *Fifth Conference on Applied Natural Language Processing*, Washington,DC, 1997. also avaliable as cmp-lg/9703004.

[Bis95] Christopher M Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[Bri90] J.S. Bridle. *Neurocomputing: Algorithms, Architectures and Applications*, chapter Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. Springer-Verlag, Berlin, 1990.

[Buø96] Finn Dag Buø. *FeasPar - A Feature Structure PARser learning to parse spontaneous speech*. PhD thesis, University of Karlsruhe, 1996.

[BW96] Finn Dag Buø and Alex Waibel. Learning to parse spontaneous speech. In *ICSLP*, 1996.

[CA97] Mark G. Core and James Allen. Coding dialogs with the damsl annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, 1997.

[Car83] Lari Carlson. *Dialogue Games: An Approach to Discourse Analysis*. D. Reidel, 1983.

[ch-97] Spanish callhome evaluation results. Hub-5 Conversational Speech Recognition Workshop, Nov. 4-6 1997.

[CII+97] 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.

[Cla96] Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.

[DH73] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[FGH+97] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal. The karlsruhe-verbmobil speech recognition engine. In *Proceedings of ICASSP 97, Muenchen, Germany*, 1997.

[Gor95] Allen Gorin. On automated language acquistition. *Journal of the Acoustical Society of America*, 97(6):3441–3461, June 1995.

[GS86] Barbara Grosz and Candace Sidner. Attention, intention and the structure of discourse. *Computational Linguistics*, 12(3):172–204, 1986.

[GZA97] Marsal Gavalda, Klaus Zechner, and Gregory Aist. High performance segmentation of spontaneous speech using part of speech and trigger word information. In *Fifth Conference on Applied Natural Language Processing*, Washington,DC, 1997.

[Hea97] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, March 1997.

[Jai91] Ajay N. Jain. *PARSEC: A Connectionists Learning Architecture for Parsing Spoken Language*. PhD thesis, School of Computer Science, Carnegie Mellon University, 1991.

[JBC+97a] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. Automatic detection of discourse structure for speech recognition and understanding. In *IEEE Workshop on Speech Recognition and Understanding*, September 1997.

[JBC+97b] Daniel Jurafsky, Rebecca Bates, Noah Coccaro, Rachel Martin, Marie Meteer, Klaus Ries, Elizabeth Shriberg, Andreas Stolcke, Paul Taylor, and Carol Van Ess-Dykema. Swbd discourse language modeling project, final report. Technical report, Johns Hopkins LVCSR Workshop-97, 1997.

[Jor95] Michael Jordan. Why the logistic function? a tutorial discussion on probabilities and neural networks. `ftp://psyche.mit.edu/pub/jordan/uai.ps.Z`, 1995.

[JSB97] Dan Jurafsky, Liz Shirberg, and Debra Biasca. Switchboard swbd-damsl coders manual. Draft 13, August 1 1997.

[KKN+97] R. Kompe, A. Kiessling, H. Nieman, E. Nöth, A. Batliner, S.Schachtl, T. Ruland, and H.U. Block. Improving parsing of spontaneous speech with the help of prosodic boundaries. In *ICASSP*, pages 811–814, Muenich, 1997.

[LM77] Joan A. Levin and Johana A. Moore. Dialogue games: Metacommunication structures for natural language interaction. *Cognitive Science*, 1(4):395–420, 1977.

[MT88] William C. Mann and Sandra Thomson. Rhetorical structure theory: Towards a functional theory of text organization. *TEXT*, 8:243–281, 1988.

[NGAH95] Christine Nakatani, Barbara J. Grosz, David D. Ahn, and Julia Hirschberg. Instructions for annotating discourses. Tr-25-95, Harvard University Center for Research in Computer Technology, Cambridge, MA, 1995.

[PL97] Rebecca J. Passonneau and Diane J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103, March 1997. 139.

[Pow79] R.J.D. Power. The organization of purposeful dialogues. *Linguistics*, 17:107–152, 1979.

[WGR97] J.H. Wright, A.L. Gorin, and G. Riccardi. Automatic acquisition of salient grammar fragments for call-type classification. In *EUROSPEECH*, Rhodes, Greece, 1997.