Proceedings of the Human Language Technology Conference, San Diego, 2001,
HLT 2001, San Diego, USA, March 18-21, 2001

Activity detection for information access to oral communication

Klaus Ries and Alex Waibel*
{ries|ahw}@cs.cmu.edu
Interactive Systems Labs, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
Interactive Systems Labs, Universität Karlsruhe, Fakultät für Informatik, 76128 Karlsruhe, Germany
http://www.is.cs.cmu.edu/ http://werner.ira.uka.de

ABSTRACT

Oral communication is ubiquitous and carries important information yet it is also time consuming to document. Given the development of storage media and networks one could just record and store a conversation for documentation. The question is, however, how an interesting information piece would be found in a large database. Traditional information retrieval techniques use a histogram of keywords as the document representation but oral communication may offer additional indices such as the time and place of the rejoinder and the attendance. An alternative index could be the activity such as discussing, planning, informing, story-telling, etc. This paper addresses the problem of the automatic detection of those activities in meeting situation and everyday rejoinders. Several extensions of this basic idea are being discussed and/or evaluated: Similar to activities one can define subsets of larger database and detect those automatically which is shown on a large database of TV shows. Emotions and other indices such as the dominance distribution of speakers might be available on the surface and could be used directly. Despite the small size of the databases used some results about the effectiveness of these indices can be obtained.

Keywords

activity, dialogue processing, oral communication, speech, information access

*We would like to thank our lab, especially Klaus Zechner, Alon Lavie and Lori Levin for their discussions and support. We would also like to thank our sponsors at DARPA. Any opinions, findings and conclusions expressed in this material are those of the authors and may not reflect the views of DARPA, or any other party.

.

1. INTRODUCTION

Information access to oral communication is becoming an interesting research area since recording, storing and transmitting large amounts of audio (and video) data is feasible today. While written information is often available electronically (especially since it is typically entered on computers) oral communication is usually only documented by constructing a new document in written form such as a transcript (court proceedings) or minutes (meetings). Oral communications are therefore a large untapped resource, especially if no corresponding written documents are available and the cost of documentation using traditional techniques is considered high: Tutorial introductions by a senior staff member might be worthwhile to attend by many newcomers, office meetings may contain informations relevant for others and should be reproducable, informal and formal group meetings may be interesting but not fully documented. In essence the written form is already a reinterpretation of the original rejoinder. Such a reinterpretation are used to

- extract and condense information
- add or delete information
- change the meaning
- cite the rejoinder
- relate rejoinders to each other

Reinterpretation is a time consuming, expensive and optional step and written documentation is combining reinterpretation and documentation step in one [1]. If however reinterpretation is not necessary or unwanted a system which is producing audiovisual records is superior. If reinterpretation is wanted or needed a system using audiovisual records may be used to improve the reinterpretation by adding all audiovisual data and the option to go back to the unaltered original. Whether reinterpretation is done or not it is crucial to be able to navigate effectively within an audiovisual document and to find a specific document.

[1]The most important exception is the literal courtroom transcript, however one could argue that even transcripts are reinterpretations since they do not contain a number of informations present in the audio channel such as emotions, hesitations, the use of slang and certain types of heteroglossia, accents and so forth. This is specifically true if transcription machines are used which restrict the transcriber to standard orthography.

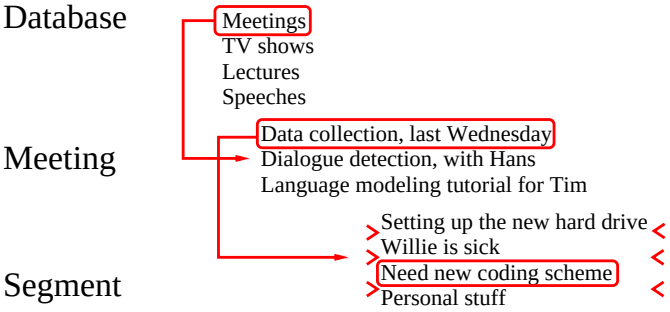

Figure 1: Information access hierarchy: Oral communications take place in very different formats and the first step in the search is to determine the database (or sub-database) of the rejoinder. The next step is to find the specific rejoinder. Since rejoinders can be very long the rejoinder has to segmented and a segment has to be selected.

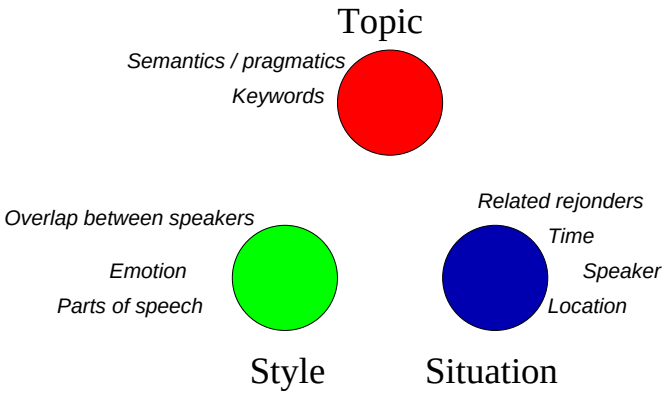

Figure 2: Bahktin's characterization of dialogue: Bahktin (1986) describes a discourse along the three major properties style, situation and topic. Current information retrieval systems focus on the topical aspect which might be crucial in written documents. Furthermore, since throughout text analysis is still a hard problem, information retrieval has mostly used keywords to characterize topic. Many features that could be extracted are therefore ignored in a traditional keyword based approach.

While keywords are commonly used in information access to written information the use of other indices such as style is still uncommon (but see Kessler et al. (1997); van Bretan et al. (1998)). Oral communication is richer than written communication since it is an interactive real time accomplishment between participants, may involve speech gestures such as the display of emotion and is situated in space and time. Bahktin (1986) characterizes a conversation by topic, situation and style. Information access to oral communication can therefore make use of indices that pertain to the oral nature of the discourse (Fig. 2). Indices other than topic (represented by keywords) increase in importance since browsing audio documents is cumbersome which makes the common interactive retrieval strategy "query, browse, reformulate" less effective. Finally the topic may not be known at all or may not be that relevant for the query formulation, for example if one just wants to be reminded what was being discussed last time a person was met. Activities are suggested as an alternative index and are a description of the type of interaction. It is common to use "action-verbs" such as story-telling, discussing, planning, informing, etc. to describe activities [2]. Items similar to activities have been shown to be directly retrievable from autobiographic memory (Herrmann, 1993) and are therefore indices that are available to participants of the conversation. Other indices may be very effective but not available: The frequency of the word "I" in the conversation, the histogram of word lengths or the histogram of pitch per participant.

In Fig. 1 the information access hierarchy is being introduced which allows to understand the problem of information access to oral communication at different levels. In Ries (1999) we have shown that the detection of general dialogue genre (database level in Fig. 1) can be done with high accuracy if a number of different example types have been annotated; in Ries et al. (2000) we have shown that it is hard but not impossible to distinguish activities in personal phone calls (segment level in Fig. 1) . In this paper we will address activities in meetings and other types of dialogues and show that these activities can be distinguished using certain features and a neural network based classifier (Sec. 2, segment level in Fig. 1). The concept of information retrieval assessment using information theoretic measures is applied to this task (Sec. 3). Additionally we will introduce a level somewhat below the database level in Fig. 1 that we call "sub-genre" and we have collected a large database of TV-shows that are automatically classified for their show-type (Sec. 4). We also explore whether there are other indices similar to activities that could be used and we are presenting results on emotions in meetings (Sec. 5).

[2] The definition of activities such as planning may vary vastly across general dialogue genres, for example compare a military combat situation with a mother child interaction. However it is often possible to develop activities and dialogue typologies for a specific dialogue genre. The related problem of general typologies of dialogues is still far from being settled and action-verbs are just one potential categorization (Fritz and Hundschnur, 1994).

2. ACTIVITY DETECTION

We are interested in the detection of activities that are described by action verbs and have annotated those in two databases:

meetings have been collected at Interactive Systems Labs at CMU (Waibel et al., 1998) and a subset of 8 meetings has been annotated. Most of the meetings are by the data annotation group itself and are fairly informal in style. The participants are often well acquainted and meet each other a lot besides their meetings.

Santa Barbara (SBC) is a corpus released by the LDC and 7 out of 12 rejoinders have been annotated.

The annotator has been instructed to segment the rejoinders into units that are coherent with respect to their topic

Activity	SBC	Meeting
Discussion	35	58
Information	25	23
Story-telling	24	10
Planning	7	19
Undetermined	5	8
Advising	5	17
Not meeting	3	2
Interrogation	2	1
Evaluation	1	0
Introduction	0	1
Closing	0	1

Table 1: Distribution of activity types: Both databases contain a lot of discussing, informing and story-telling activities however the meeting data contains a lot more planning and advising.

and activity and annotate them with an activity which follows the intuitive definition of the action-verb such as discussing, planning, etc. Additionally an activity annotation manual containing more specific instructions has been available (Ries et al., 2000; Thymé-Gobbel et al., 2001) [3]. The list of tags and the distribution can be seen in Tab. 1. The set of activities can be clustered into "interactive" activities of equal contribution rights (discussion,planning), one person being active (advising, information giving, story-telling), interrogations and all others.

Measure	Meeting		SBC		CallHome
	all	inter	all	inter	Spanish
κ	0.41	0.51	0.49	0.56	0.59
Mutual inf.	0.35	0.25	0.65	0.32	0.61

Table 2: Intercoder agreement for activities: The meeting dialogues and Santa Barbara corpus have been annotated by a semi-naive coder and the first author of the paper. The κ-coefficient is determined as in Carletta et al. (1997) and mutual information measures how much one label "informs" the other (see Sec. 3). For CallHome Spanish 3 dialogues were coded for activities by two coders and the result seems to indicate that the task was easier.

Both datasets have been annotated not only by a semi-naive annotator but also by the first author of the paper. The results for κ-statistics (Carletta et al., 1997) and mutual information between the coders can be seen in Tab. 2. The intercoder agreement would be considered moderate but compares approximately to Carletta et al. (1997) agreement on transactions ($\kappa = 0.59$), especially for the interactive activities and CallHome Spanish.

For classification a neural network was trained that uses the softmax function as its output and KL-divergence as the error function. The network connects the input directly to the output units. Hidden units have not been used since they did not yield improvements on this task. The network was trained using RPROP with momentum (Riedmiller and Braun, 1993) and corresponds to an exponential model (Nigam et al., 1999). The momentum term can be interpreted as a Gaussian prior with zero mean on the network weights. It is the same architecture that we used previously (Ries et al., 2000) for the detection of activities on CallHome Spanish. Although some feature sets could be trained using the iterative scaling algorithm if no hidden units are being used the training times weren't high enough to justify the use of the less flexible iterative scaling algorithm. The features used for classification are

Feature	all		interactive	
	SBC	meet	SBC	meet
baseline	32.7	41.1	50.5	54.6
dialogue acts per channel	28.1	37.6	47.7	56.7
dialogue acts	28.0	36.2	46.7	65.3
words	38.3	39.7	53.3	54.6
dominance	32.7	44.7	64.5	58.2
style	24.3	35.5	53.3	58.9
style + words	42.1	38.3	52.3	57.5
dominance + words	41.1	41.1	52.3	58.9
dominance + style + words	42.1	39.7	53.3	60.3
dialogue acts + words	42.1	37.6	57.0	61.0
dialogue acts + style + words	39.3	40.4	57.9	61.0
Wordnet	37.4	37.6	46.7	52.5
Wordnet + words	49.5	39.0	53.3	57.5
first author	59.8	57.9	73.8	72.7

Table 3: Activity detection: Activities are detected on the Santa Barbara Corpus (SBC) and the meeting database (meet) either without clustering the activities (all) or clustering them according to their interactivity (interactive) (see Sec. 2 for details).

words the 50 most frequent words / part of speech pairs are used directly, all other pairs are replaced by their part of speech [4].

stylistic features adapted from Biber (1988) and contain mostly syntactic constructions and some word classes.

Wordnet a total of 40 verb and noun classes (so called lexicographers classes (Fellbaum, 1998)) are defined and a word is replaced by the most frequent class over all possible meanings of the word.

dialogue acts such as statements, questions, backchannels, ... are detected using a language model based detector trained on Switchboard similar to Stolcke et al. (2000) [5]

[3] In contrast to (Ries et al., 2000; Thymé-Gobbel et al., 2001) the "consoling" activity has been eliminated and an "informing" activity has been introduced for segments where one or more than one member of the rejoinder give information to the others. Additionally an "introducing" activity was added to account for a introduction of people or topics at the beginning of meetings.

[4] Klaus Zechner trained an English part of speech tagger tagger on Switchboard that has been used. The tagger uses the code by Brill (1994).

[5] The model was trained to be very portable and therefore the following choices were taken: (a) the dialogue model is context-independent and (b) only the part of speech are taken as the input to the model plus the 50 most likely word/part of speech types.

dominance is described as the distribution of the speaker dominance in a conversation. The distribution is represented as a histogram and speaker dominance is measured as the average dominance of the dialogue acts (Linell et al., 1988) of each speaker. The dialogue acts are detected and the dominance is a numeric value assigned for each dialogue act type. Dialogue act types that restrict the options of the conversation partners have high dominance (questions), dialogue acts that signal understanding (backchannels) carry low dominance.

First author The activities used for classification are those of the semi-naive coder. The "first author" column describes the "accuracy" of the first author with respect to the naive coder.

The detection of interactive activities works fairly well using the dominance feature on SBC which is also natural since the relative dominance of speakers should describe what kind of interaction is exhibited. The dialogue act distribution on the other hand works fairly well on the more homogeneous meeting database were there is a better chance to see generalizations from more specific dialogue based information. Overall the combination of more than one feature is really important since word level, Wordnet and stylistic information, while sometimes successful, seem to be able to improve the result while they don't provide good features by themselves. The meeting data is also more difficult which might be due to its informal style.

3. INFORMATION ACCESS ASSESSMENT

Assuming a probabilistic information retrieval model a query r – in our example an activity – predicts a document d with the probability $q(d|r) = \frac{q(r|d)q(d)}{q(r)}$. Let $p(d,r)$ be the real probability mass distribution of these quantities. The probability mass function $q(r|d)$ is estimated on a separate training set by a neural network based classifier [6]. The quantity we are interested in is the reduction in expected coding length of the document using the neural network based detector [7]:

$$-E_p \log \frac{q(D)}{q(D|R)} \approx H(R) - E_p \log \frac{1}{q(R|D)}$$

The two expectations correspond exactly to the measures in Tab. 5, the first represents the baseline, the second the one for the respective classifier. In more standard information theoretic notation this quantity may be written as:

$$H(R) - (H_p(R|D) + D(p(r|d)||q(r|d)))$$

This equivalence is not extremely useful though since the quantities in parenthesis can't be estimated separately. For the small meeting database and SBC however no entropy reductions could be obtained. On the larger databases, on the other hand, entropy reductions could be obtained ($\approx$ 0.5bit on the CallHome Spanish database Ries et al. (2000), $\approx$ 1bit for the sub-database detection problem in Sec. 4).

[6]All quantities involving the neural net $q(r|d)$ have been determined using a round robin approach such that network is trained on a separate training set.

[7]Since estimating $q(d)$ is simple we may assume that $q(d) \approx \sum_r p(d,r)$.

Another option is to assume that the labels of one coder are part of D. If the query by the other coder is R we are interested in the reduction of the document entropy given the query. If we furthermore assume that $H(R|D) = H(R|R')$ where R' is the activity label embedded in D:

$$H(D) - H(D|R) = H(R) - H(R|D) = MI(R, R')$$

Tab. 2 shows that the labels of the semi-naive coder and the first author only inform each other by $0.25 - 0.65$ bits. However, since all constraints are important to apply, it might be important to include manual annotations to be matched by a query or in a graphical presentation of the output results.

Another interesting question to consider is whether the activity is correlated with the rejoinder or not. This question is important since a correlation of the activity with the rejoinder would mean that the indexing performance of activities needs to be compared to other indices that apply to rejoinders such as attendance, time and place (for results on the correlation with rejoinders see Waibel et al. (2001)). The correlation can be measured using the mutual information between the activity and the meeting identity. The mutual information is moderate for SBC ($\approx$ 0.67 bit) and much lower for the meetings ($\approx$ 0.20 bit). This also corresponds to our intuition since some of the rejoinders in SBC belong to very distinct dialogue genre while the meeting database is homogeneous. The conclusion is that activities are useful for navigation in a rejoinder if the database is homogeneous and they might be useful for finding conversations in a more heterogeneous database.

	#		#		#
Talk	344	Edu	25	Finance	8
News	217	Scifi	24	Religious	5
Sitcom	97	Series	24	Series-Old	3
Soap	87	Cartoon	23	Infotain	3
Game	46	Movies	22	Music	2
Law	32	Crafts	17	Horror	1
Sports	32	Specials	15		
Drama	31	Comedy	9		

Table 4: TV show types: The distribution of show types in a large database of TV shows (1067 shows) that has been recorded over the period of a couple of months until April 2000 in Pittsburgh, PA

4. DETECTION OF SUB-DATABASES

We set up an environment for TV shows that records the subtitles with timestamps continuously from one TV channel and the channel was switched every other day. At the same time the TV program was downloaded from http://tv.yahoo.com/ to obtain programming information including the genre of the show. Yahoo assigns primary and secondary show types and unless the combination of primary/secondary show-type is frequent enough the primary showtype is used (Tab. 4). The TV show database has the advantage that we were able to collect a large and varied database with little effort. The same classifier as in Sec. 2 has been used however dialogue acts have not been detected since the data contains a lot of noise, is not necessarily conversational and speaker identities can't be determined easily. Detection results for TV shows can be seen in Tab. 5. It may

be noted that adding a lot of keywords does improve the detection result but not so much the entropy. It may therefore be assume that there is a limited dependence between topic and genre which isn't really a surprise since there are many shows with weekly sequels and there may be some true repeats.

Feature			accuracy	entropy
Wordnet	stylistic	words		
	baseline		32.2	3.31
	•		50.9	2.73
	•	50	62.2	2.33
•	•	50	60.0	2.29
•	•		61.2	2.28
•			56.9	2.41
•		50	61.5	2.25
		50	61.3	2.35
		250	62.7	2.17
		500	66.0	2.14
•	•	500	64.9	2.13
		5000	67.2	2.08

Table 5: Show type detection: Using the neural network described in Sec. 2 the show type was detected. If there is a number in the word column the word feature is being used. The number indicates how many word/part of speech pairs are in the vocabulary additionally to the parts of speech.

5. EMOTION AND DOMINANCE

Emotions are displayed in a variety of gestures, some of which are oral and may be detected via automated methods from the audio channel (Polzin, 1999). Using only verbal information the emotions happy, excited and neutral can be detected on the meeting database with 88.1% accuracy while always picking neutral yields 83.6%. This result can be improved to 88.6% by adding pitch and power information.

While these experiments were conducted at the utterance level emotions can be extended to topical segments. For that purpose the emotions of the individual utterances are entered in a histogram over the segment and the vectors are clustered automatically. The resulting clusters roughly correspond to a "neutral", "a little happy" and "somewhat excited" segment. Using the classifier for emotions on the word level the segment can be classified automatically into categories with a 83.3% accuracy while the baseline is 68.9%. The entropy reduction by automatically detected emotional activities is ≈ 0.3bit [8]. A similar attempt can be made for dominance (Linell et al., 1988) distributions: Dominance is easy to understand for the user of an information access system and it can be determined automatically with high accuracy.

[8] A similar classification result for emotions on the utterance level has been obtained by just using the laughter vs. non-laughter tokens of the transcript as the input. This may indicate that (a) the index should really be the amount of laughter in the conversational segment and that (b) emotions might not be displayed very overtly in meetings. These results however would require a wider sampling of meeting types to be generally acceptable.

6. CONCLUSION AND FUTURE WORK

It has been shown that activities can be detected and that they may be efficient indices for access to oral communication. Overall it is easy to make high level distinctions with automated methods while fine-grained distinctions are even hard to make for humans – on the other hand automatic methods are still able to model some aspect of it (Fig. 3). To obtain an reduction in entropy a relatively large database such as CallHome Spanish is required (120 dialogues). Alternatives to activities might be emotional and dominance distributions that are easier to detect and that may be natural to understand for users. If activities are only used for local navigation support within a rejoinder one could also visualize by displaying the dialogue act patterns for each channel on a time line.

The author has also observed that topic clusters and activities are largely independent in the meeting domain resulting in orthogonal indices. Since activities have intuitions for naive users and they may be remembered it can be assumed that users would be able to make use of these constraints. Ongoing work includes the use of speaker activity for dialogue segmentation and further assessment of features for information access. Overall the methods presented here and the ongoing work are improving the ability to index oral communication. It should be noted that some of the techniques presented lend themselves to implementations that don't require (full) speech recognition: Speaker identification and dialogue act identification may be done without an LVCSR system which would allow to lower the computational requirements as well as to a more robust system.

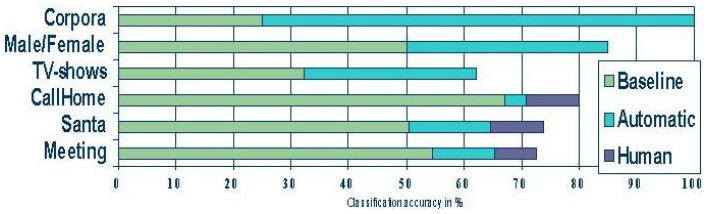

Figure 3: Detection accuracy summary: The detection of high-level genre as exemplified by the differentiation of corpora can be done with high accuracy using simple features (Ries, 1999). Similar it was fairly easy to discriminate between male and female speakers on Switchboard (Ries, 1999). Discriminating between sub-genre such as TV-show types (Sec. 4) can be done with reasonable accuracy. However it is a lot harder to discriminate between activities within one conversation for personal phone calls (CallHome) (Ries et al., 2000) or for general rejoinders (Santa) and meetings (Sec. 2).

References

M. M. Bahktin. *Speech Genres and other late Essays*, chapter Speech Genres. University of Texas Press, Austin, 1986.

D. Biber. *Variation across speech and writing*. Cambridge University Press, 1988.

E. Brill. A report on recent progress in transformation based error-driven learning. In *DARPA Workshop*, 1994.

J. Carletta, A. Isard, S. Isard, J. C. Kowtko, G. Doherty-Sneddon, and A. H. Anderson. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31, March 1997.

C. Fellbaum, editor. *WordNet – An Electronic Lexical Database*. MIT press, 1998.

G. Fritz and F. Hundschnur. *Handbuch der Dialoganalyse*. Niemeyer, Tuebingen, 1994.

D. J. Herrmann. *Autobiographical memory and the validity of retrospective reports*, chapter The validity of retrospective reports as a function of the directness of retrieval processes, pages 21–31. Springer, 1993.

B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 32–38. Morgan Kaufmann Publishers, San Francisco CA, 1997. URL http://xxx.lanl.gov/abs/cmp-lg/9707002.

P. Linell, L. Gustavsson, and P. Juvonen. Interactional dominance in dyadic communication: a presentation of initiative-response analysis. *Linguistics*, 26:415–442, 1988.

K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999. URL http://www.cs.cmu.edu/~lafferty/.

T. Polzin. *Detecting Verbal and Non-Verbal Cues in the Communication of Emotion*. PhD thesis, Carnegie Mellon University, November 1999.

M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proc. of the IEEE Int. Conf. on Neural Networks*, pages 586–591, 1993.

K. Ries. Towards the detection and description of textual meaning indicators in spontaneous conversations. In *Proceedings of the Eurospeech*, volume 3, pages 1415–1418, Budapest, Hungary, September 1999.

K. Ries, L. Levin, L. Valle, A. Lavie, and A. Waibel. Shallow discourse genre annotation in callhome spanish. In *Proceecings of the International Conference on Language Ressources and Evaluation (LREC-2000)*, Athens, Greece, May 2000.

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

A. Thymé-Gobbel, L. Levin, K. Ries, and L. Valle. Dialogue act, dialogue game, and activity tagging manual for spanish conversational speech. Technical report, Carnegie Mellon University, 2001. in preperation.

van Bretan, J. Dewe, A. Hallberg, J. Karlgren, and N. Wolkert. Genres defined for a purpose, fast clustering, and an iterative information retrieval interface. In *Eighth DELOS Workshop on User Interfaces in Digital Libraries Långholmen*, pages 60–66, October 1998.

A. Waibel, M. Bett, and M. Finke. Meeting browser: Tracking and summarising meetings. In *Proceedings of the DARPA Broadcast News Workshop*, 1998.

A. Waibel, M. Bett, F. Metze, K. Ries, T. Schaaf, T. Schultz, H. Soltau, H. Yu, and K. Zechner. Advances in automatic meeting record creation and access. In *ICASSP*, Salt Lake City, Utah, USA, 2001. to appear.