Discourse Language-Modeling Project

Rebecca Bates	*Boston University*
Noah Coccaro	*University of Colorado*
Carol Van Ess-Dykema	*DoD*
Dan Jurafsky	*University of Colorado*
Rachel Martin	*Johns Hopkins University*
Marie Meteer	*BBN*
Klaus Ries	*CMU/Universität Karlsruhe*
Liz Shriberg	*SRI*
Andreas Stolcke	*SRI*
Paul Taylor	*University of Edinburgh*

November 5, 1997

Johns Hopkins LVCSR Workshop 97

Overview of Talk

- Summer Workshop main ideas
- Model Description
- Results
- Other work on the workshop & related work
- Conclusions

Goal: Better Use of Discourse Knowledge

Understanding Spontaneous Dialog:

A meeting summarizer needs to know who said what to who.

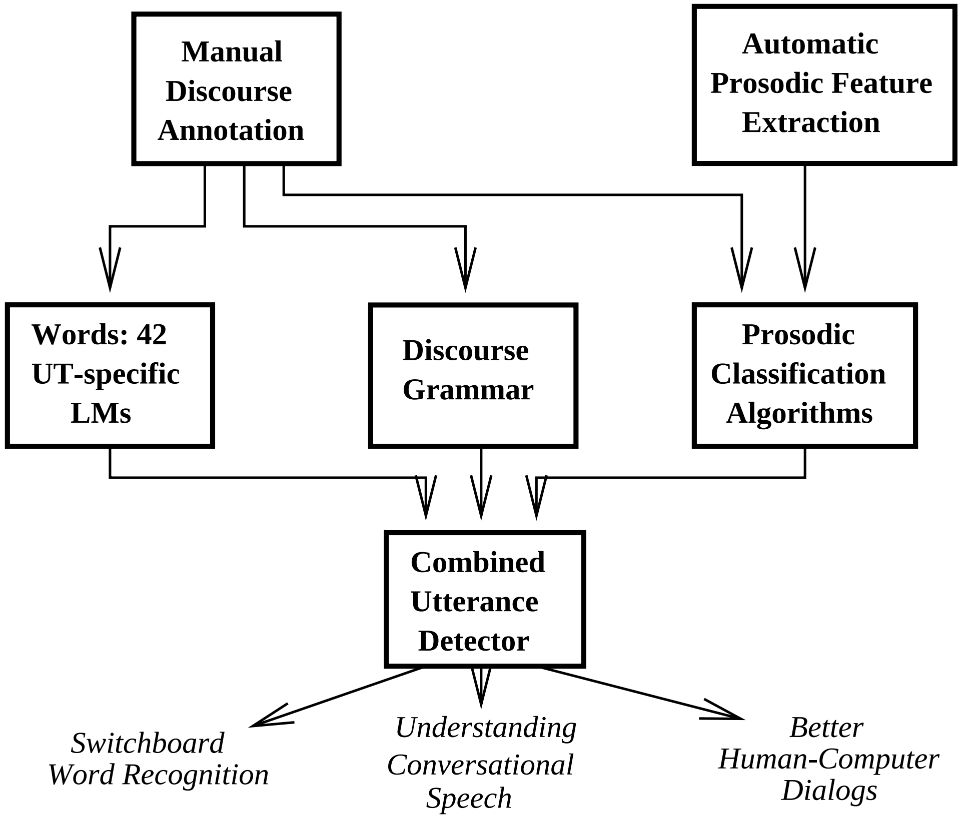
Better Human-Computer Dialog:

A conversational agent needs to know whether it was just asked a question or ordered to do something.

Improved Switchboard Word Recognition:

Different utterance types have very different words.

Logistics


Manual
Discourse
Annotation
Automatic
Prosodic Feature
Extraction
Words: 42
UT-specific
LMs
Discourse
Grammar
Prosodic
Classification
Algorithms
Combined
Utterance
Detector
Switchboard
Word Recognition
Understanding
Conversational
Speech
Better
Human-Computer
Dialogs

Speech Act Coding Scheme

- ≈ 60 partially orthogonal tags
- coders used 220 tag combinations, clustered into 42 tags for the WS
- compatible with DAMSL
- 85% of the words are in statements $\leadsto$ little effect on WE
- intercoder agreement high $\kappa = 0.8$
- average conversation
  - 144 turns
  - 271 utterances
  - takes 28 minutes to label (at the end of the project)

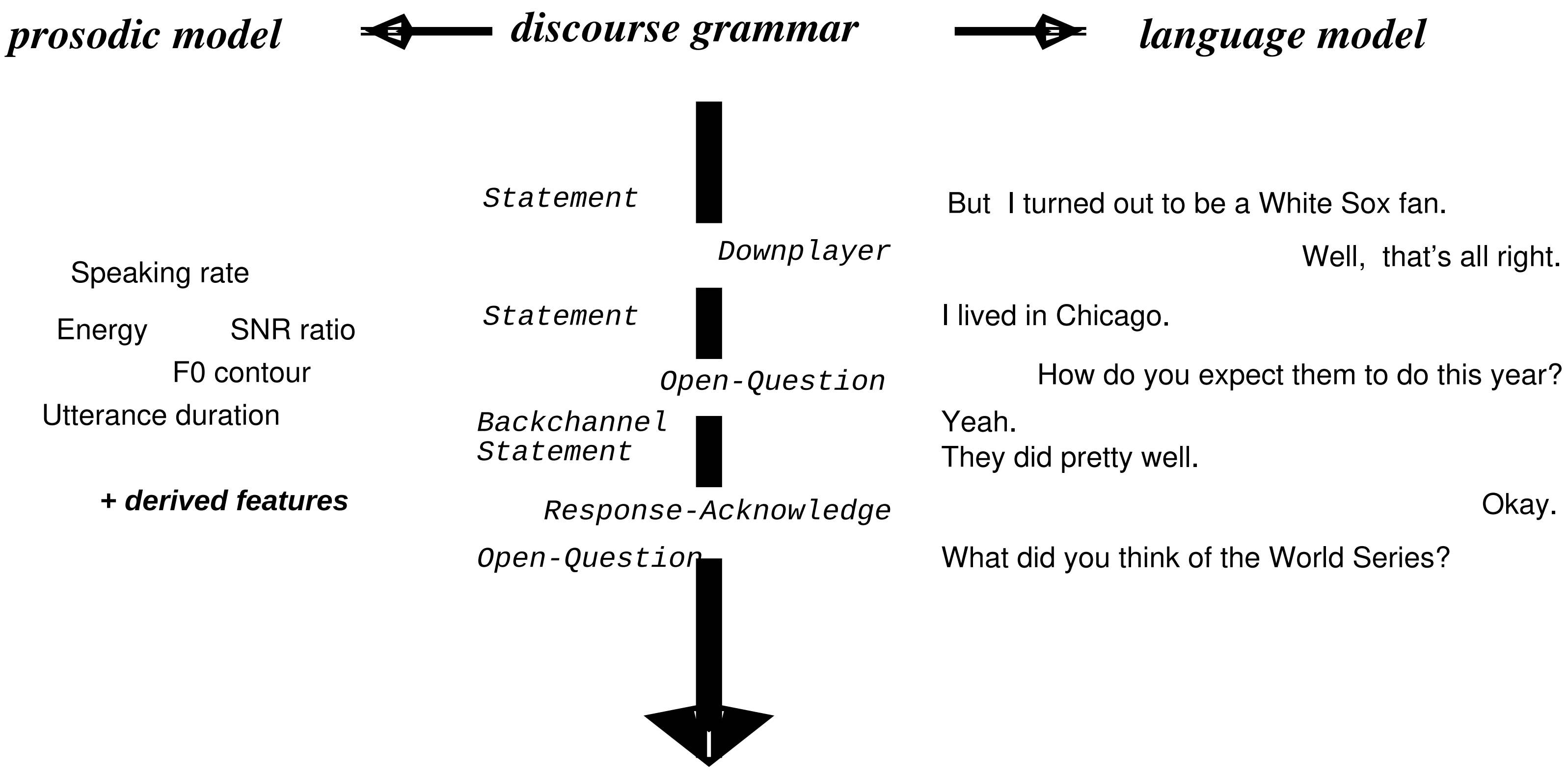
Speech Act Coding Scheme

Tag	Example	Cnt	%
Statement-non-opinion	*Me, I'm in the legal department.*	72,824	36%
Backchannel	*Uh-huh.*	37,096	19%
Statement-opinion	*I think it's great*	25,197	13%
Agree/Accept	*That's exactly it.*	10,820	5%
Abandoned or Turn-Exit	*So, -/*	10,569	5%
Appreciation	*I can imagine.*	4,633	2%
Yes-No-Question	*Do you have to have any special training?*	4,624	2%
Non-verbal	*<Laughter>,<Throat_clearing>*	3,548	2%
Yes answers	*Yes.*	2,934	1%
Conventional-closing	*Well, it's been nice talking to you.*	2,486	1%
Uninterpretable	*But, uh, yeah*	2,158	1%
Wh-Question	*Well, how old are you?*	1,911	1%
No answers	*No.*	1,340	1%
Response Ack	*Oh, okay.*	1,277	1%
Hedge	*I don't know if I'm making any sense or not.*	1,182	1%
Declarative Question	*So you can afford to get a house?*	1,174	1%
Other	*Well give me a break, you know.*	1,074	1%
Backchannel-Question	*Is that right?*	1,019	1%

What is Discourse Structure? An Excerpt

A	**B**	A *B*
Wh-Question		What kind do you have now?
	Statement	*Uh, we have a, a Mazda nine twenty nine and a Ford Crown Victoria and a little two seater CRX.*
Acknowledge-Answer		Oh, okay.
	Opinion	*Uh, it's rather difficult to, to project what kind of, uh, -*
Statement		we'd, look, always look into, uh, consumer reports to see what kind of, uh, report, or, uh, repair records that the various cars have –
. . .		
	Turn-Exit	*So, uh, -*
Yes-No-Quest		And did you find that you like the foreign cars better than the domestic?
	Answer-Yes	*Uh, yeah,*
. . .		
	Statement	*We've been extremely pleased with our Mazdas.*
Backchannel-Quest		Oh, really?
	Answer-Yes	*Yeah.*

Overall Model



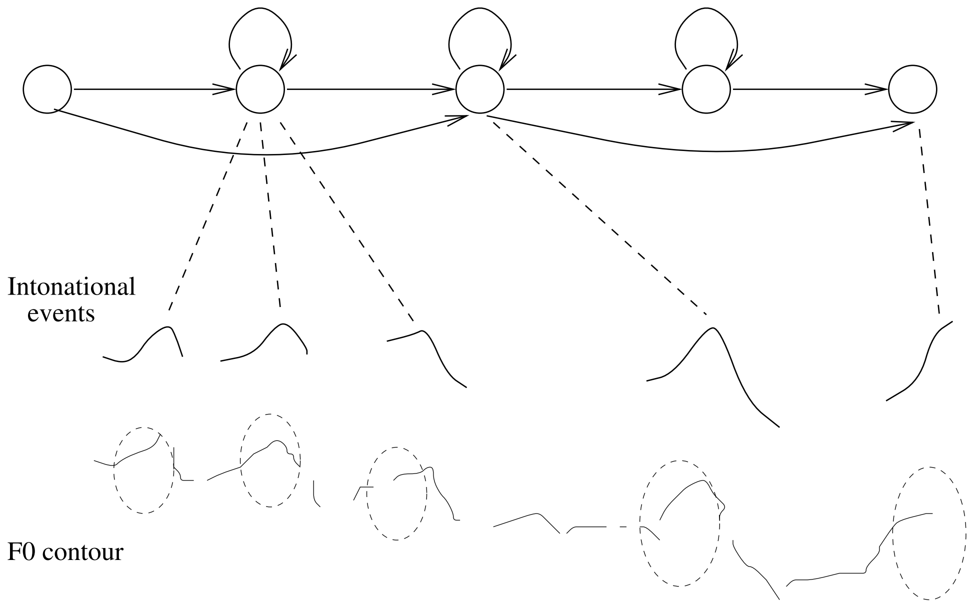
Discourse Grammar

- simple backoff ngram model works fine
- important to encode channel information
  ($p(statement|statement/same\_channel)$)
- long range & exponential models unsuccessful
- $\leadsto$ local coordination of utterance types

Speech Act Specific Grammars

- build 42 speech act specific grammars
- cheating experiment: 1% absolute WE improvement
- strongest constraint for speech act detection
- together with unigram discourse grammar: close to best result

Prosody – Hidden Markov Contour Model



- detect intonational events from pitch contour
- detect utterance types from intonational events
- reflects linguistic intuitions about intonation

Prosody – Decision Trees

- construct a feature vector of prosodic features
- trajectories categorized by means, ratios, variances and simple regressions
- train classifier (decision tree) for speech acts on feature vectors
- reasonable good performance alone
- after workshop: also tried neural nets and exponential models

Integration

Maximize the posterior probability of the utterance types:

$$
\begin{aligned}
U^* &= \max_U P(U|W,F) \\
&= \max_U P(U)P(W,F|U) \\
&\approx \max_U P(U)P(W|U)P(F|U) \\
&\approx \max_U P(U)\prod_i P(W_i|U_i)P(F_i|U_i)
\end{aligned}
$$

To use prosodic classifier $P(F_i|U_i) = P(U_i|F_i)P(F_i)/P(U_i) \propto P(U_i|F_i)/P(U_i)$

Detection Results

Model	Accuracy (%)
Chance	**35**
Prosodic (HMM) + Discourse Unigrams	**42**
Prosodic (Decision Tree) + Discourse Bigrams	**50**
Words (1-best-list) + Discourse Bigrams	**63.5**
Words (2500-best-list) + Discourse Bigrams	**64.6**
Prosodic (DT) + Words (2500) + Discourse Bigrams	**65.0**
Words (Cheating correct) + Discourse Bigrams	**71.0**
Human Labelers	**83**

Cheating Results

Utterance Type	Word Error (%) Baseline	Oracle	Δ
Answer No	29.4	11.8	-17.6
Backchannel	25.9	18.6	-7.3
Backchannel Questions	15.2	9.1	-6.1
Abandoned & Turn-Exit	48.9	45.2	-3.7
Wh-Questions	38.4	34.9	-3.5
Yes-No-Questions	55.5	52.3	-3.2
Statement	42.0	41.5	-0.5
all	41.2	40.3	-0.9

Recognition Results

Cheating LM: Condition LM on true UT
1-best LM: Condition on the 1-best guess of UT
Mixture LM: Mixture of LMs using posterior UT probabilities

Model	Word Error (%)	Perplexity
Baseline	41.2	76.8
1-best LM	41.0	69.8
Mixture LM	40.9	66.9
Cheating LM	40.3	66.8

Note: Mixture models requires full word history for evaluation, efficient lattice rescoring not possible. Used 2500-best list which gave same WER as lattice for baseline and cheating model.

Erroranalysis - “linguistically engineered”

manual	semi-automatic	automatic
"language experts"	"linguistic engineering"	"engineers"
large amount of data (e.g. transcript reading)	moderate amount of data (e.g. selected data reading. vizualization)	small amount of data (e.g. word accuracy number)

- provided basic results throughout the WS
- erroranalysis tool for selective transcript reading
- developed visualization techniques
- other examples: salience analysis, utterance initial analysis, ...

Above & Below the Statement

Divide statements into contextdependent groups:

Answers	Short
Following another person's Question Marked sd^e (extended answers)	Following Another person's Statement Following own backchannel

Long	Medium
Preceding another person's Backchannel Preceding or following own sv	Following own sd Other

Looking into the statement:

- **Given/new**: Sentences are not uniform from beginning to end
- Given information tends to occur early
- New information tends to occur late
- we can define a pivot based on the verb

Related Work

VerbMobil various similar approaches, including segmentation

Segmentation work by Stolcke/Shriberg, Gavalda/Zechner/Aist, Buø on automated segmentation

Taylor similar yet simpler model slightly better on MapTask

Gorin call type classification is a *very* similar problem

- uses grammar inference
- uses a more direct classification approach
- no integration of prosody

Clarity DoD sponsored project on discourse processing of CALLHOME SPANISH

After Workshop Activities

- continuation of prosody work
- revisit independence assumptions
- debug experiments and open ends
- publications
  - ASRU workshop paper (8 page) – final version done
  - WS report ($\approx$ 30-40 pages) – done
  - Prosodic features – special issue in *Language and Speech*, submitted
  - AAAI98 spring symposium on Applying Machine Learning to Discourse Processing – submitted
  - and many more planned

Conclusions

- established a tagging scheme for speech acts for SWBD
- provided a database for doing research
- proved the feasibility of speech act tagging on SWBD
- provided simple baseline models on a number of problems
- gave interesting insights into the data
- provided know-how for approaching these problems in the future