Johns Hopkins LVCSR Workshop 97
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*

Final Report

Revised, October 16, 1997

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.*

Goal: Better Use of Discourse Knowledge

Why should we care?

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 just got asked a question or ordered to do something.

Improved Switchboard Word Recognition:

Different Utterance types have very different words.

Why Discourse Knowledge for Speech?

- word **do** has error rate of 72%.
- **do** is in almost every **Yes-No-Question**
- If detect **Yes-No-Question** could increase **P(do)**
- Use rising intonation to detect **Yes-No-Question**

```
CORRECT:    um DO YOU kind of think it's something ELSE
RECOGNIZED: um ** THE kind of think it's something ****

CORRECT:    DO you go see MU- a lot of ***** MUSICALS
RECOGNIZED: ** you go see *** a lot of MUSIC ALL
```

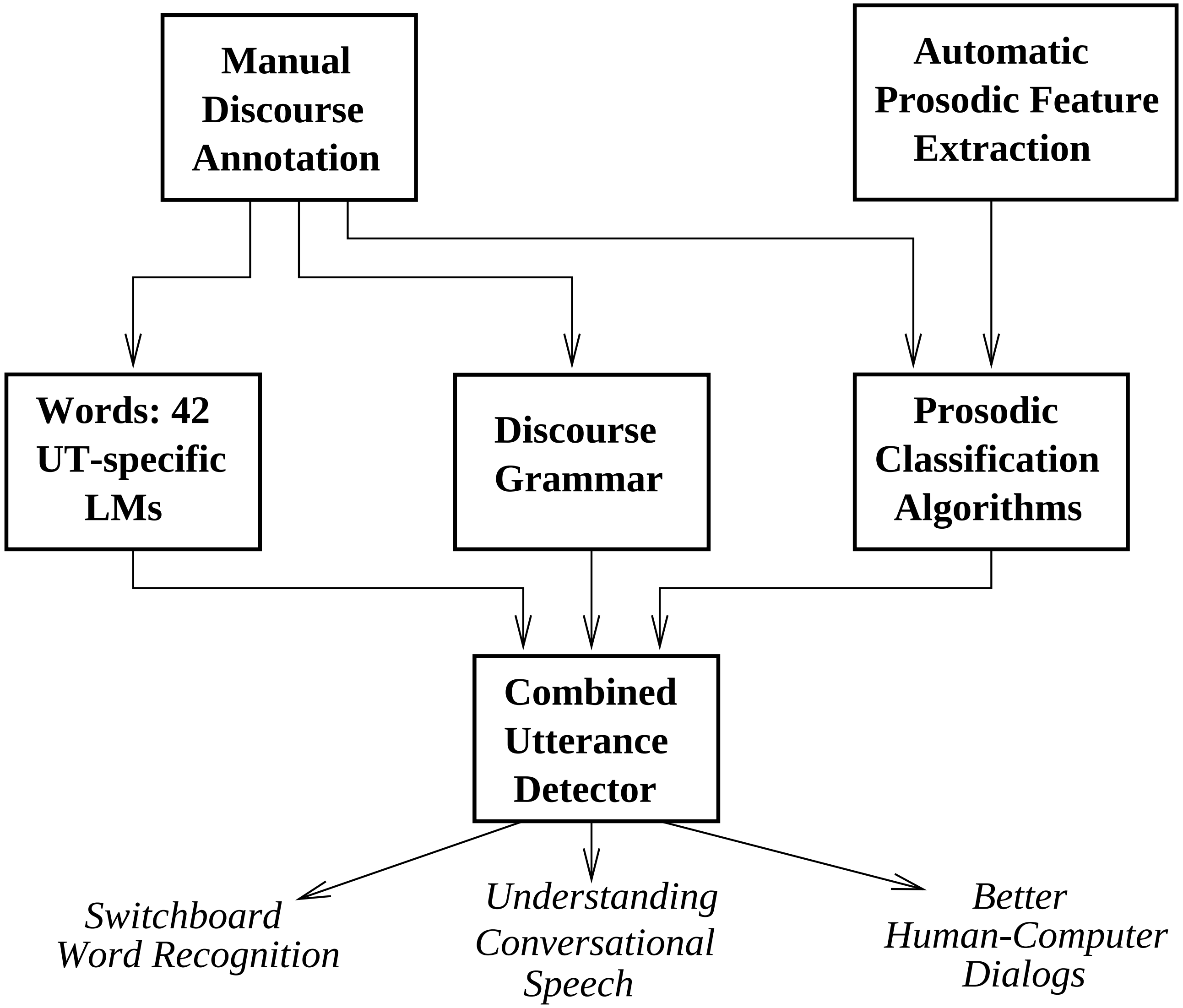
Summer 1997 Summary

Data Annotation: Annotate 1155 conversations (1.4 million words) with hand-labeled discourse-tags and automatically extracted prosodic features

Detect Utterance Type: Train various automatic utterance-type detectors

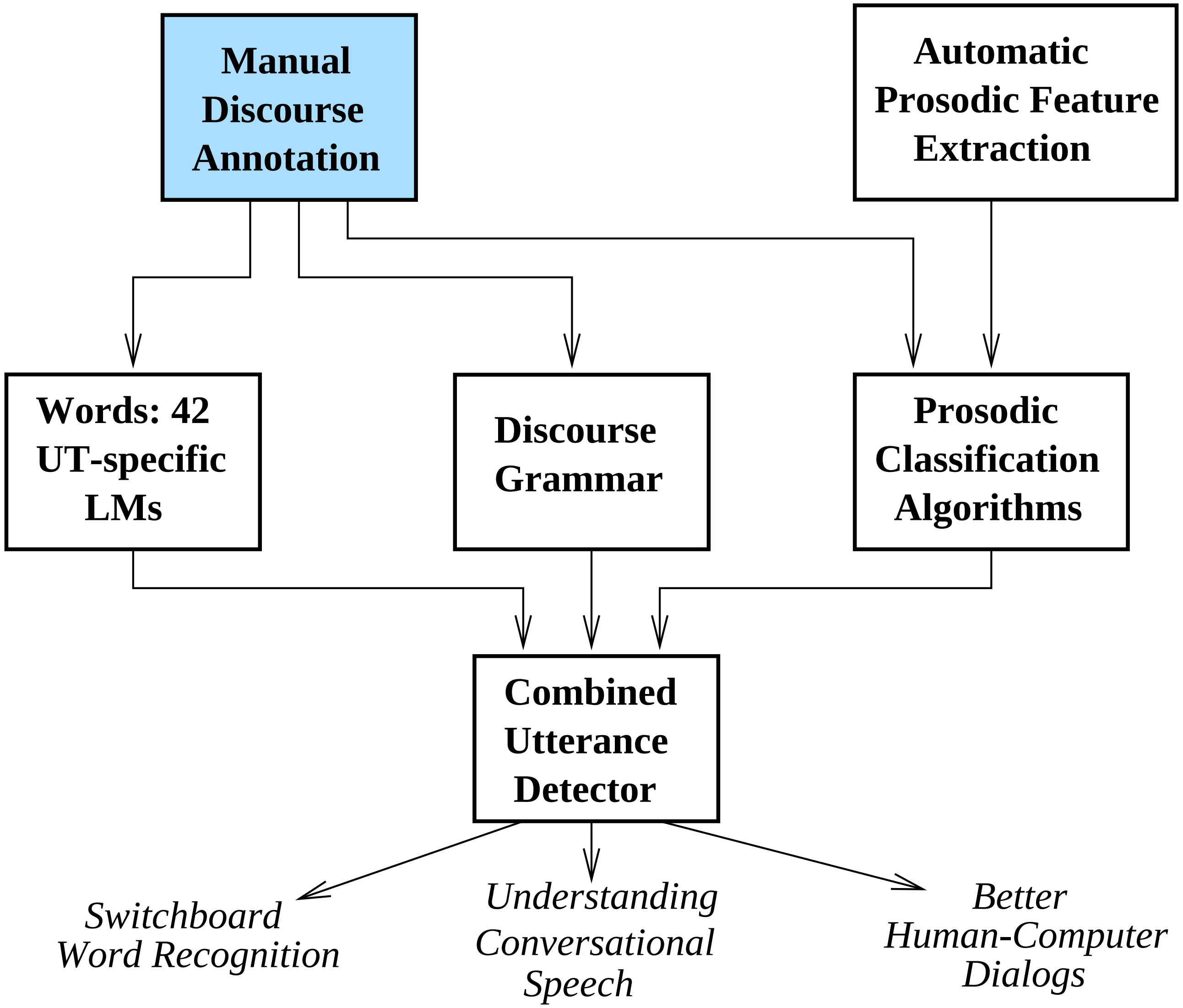
Decrease Switchboard Word Error (WER): Use utterance-type to improve Switchboard (SWBD) Language Models (LMs)

Outline of Talk and Project



Part I: Manual Discourse Tagging

Outline of Talk and Project



Discourse-Tagging 1155 SWBD Conversations

1155 conversations: our 'SWBD-DAMSL' tagset, extension of DAMSL tag-set (Allen and Core 1997)

Many Thanks: to Susann LuperFoy, Nigel Ward, James Allen, Julia Hirschberg, Marilyn Walker for tag-set advice

1.4 million words: order of magnitude larger than other discourse databases

Other tagsets: task-oriented or human-computer dialogs (Carletta et al 1995, Reithinger 1995, etc)

~60 groupable labels $\rightarrow$ 220 tags $\rightarrow$ **42** final tags

each utterance: receives one label (utterance $\approx$ sentence)

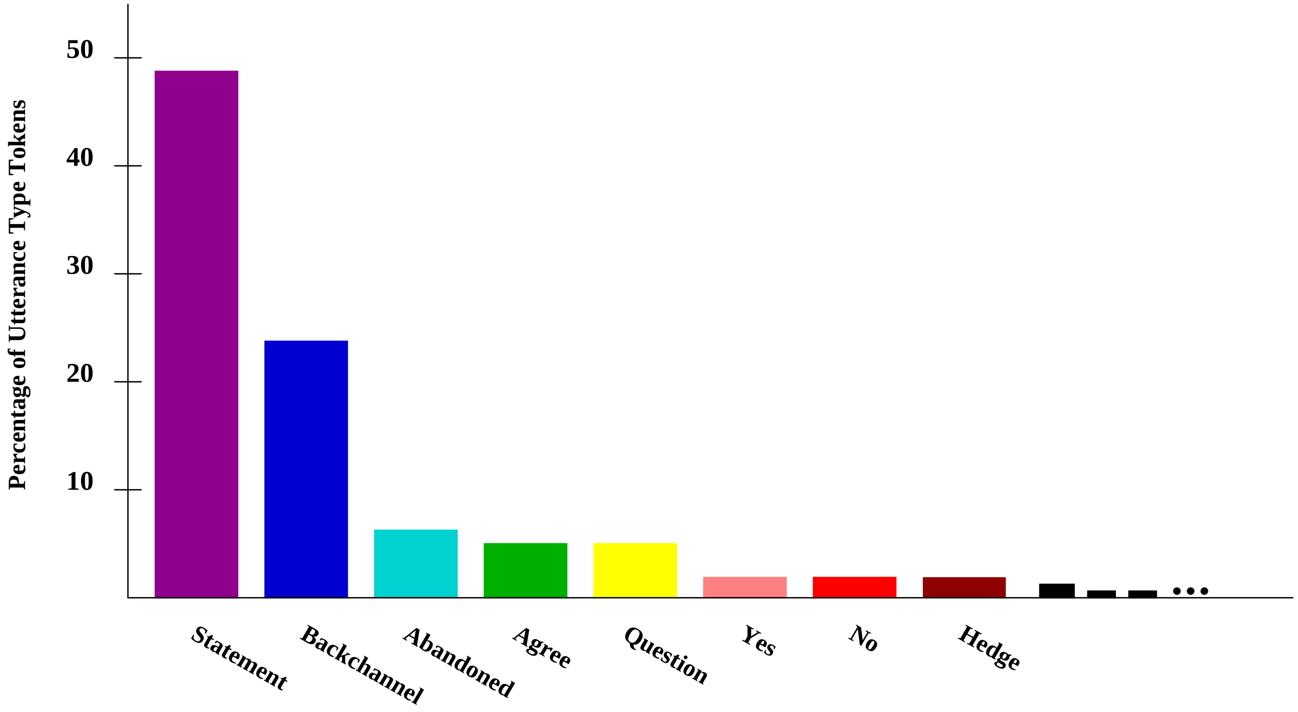
18 most frequent tags (of 42)

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%

Discourse Tagging Logistics

- 7 labelers are Boulder linguistics grad students: Debra Biasca, Marion Bond, Traci Curl, Anu Erringer, Michelle Gregory, Lori Heintzelman, Taimi Metzler, Amma Oduro
- Average conversation:
  - 144 turns
  - 271 utterances
  - takes 28 minutes to label

Distribution of Utterance Types



Error Rates on Specific Utterance Types

Utterance Type	WER	Oracle WER	Improvement with Oracle
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%

Examples: Questions

Yes-No-Question	*Do you have to have any special training?*
Yes-No-Question	*But that doesn't eliminate it, does it?*
Yes-No-Question	*Uh, I guess a year ago you're probably* watching C N N a lot, right?
Declarative-Question	*So you're taking a government course?*
Wh-Question	*Well, how old are you?*

Examples: Backchannels

A	**B**	A	*B*
	Statement		*but, uh, we're to the point now where our financial income is enough that we can consider putting some away –*
Backchannel		Uh-huh. /	
	Statement		*– for college, /*
	Statement		*so we are going to be starting a regular payroll deduction –*
Backchannel		Um. /	
	Statement		*– in the fall /*
	Statement		*and then the money that I will be making this summer we'll be putting away for the college fund. /*
Appreciation		Um. Sounds good. /	

Examples: Statements

Statement	*Well, we have a cat, um,*
Statement	*He's probably, oh, a good two years old, big, old, fat and sassy tabby.*
Statement	*He's about five months old*
Opinion	*Well, rabbits are darling.*
Opinion	*I think it would be kind of stressful.*

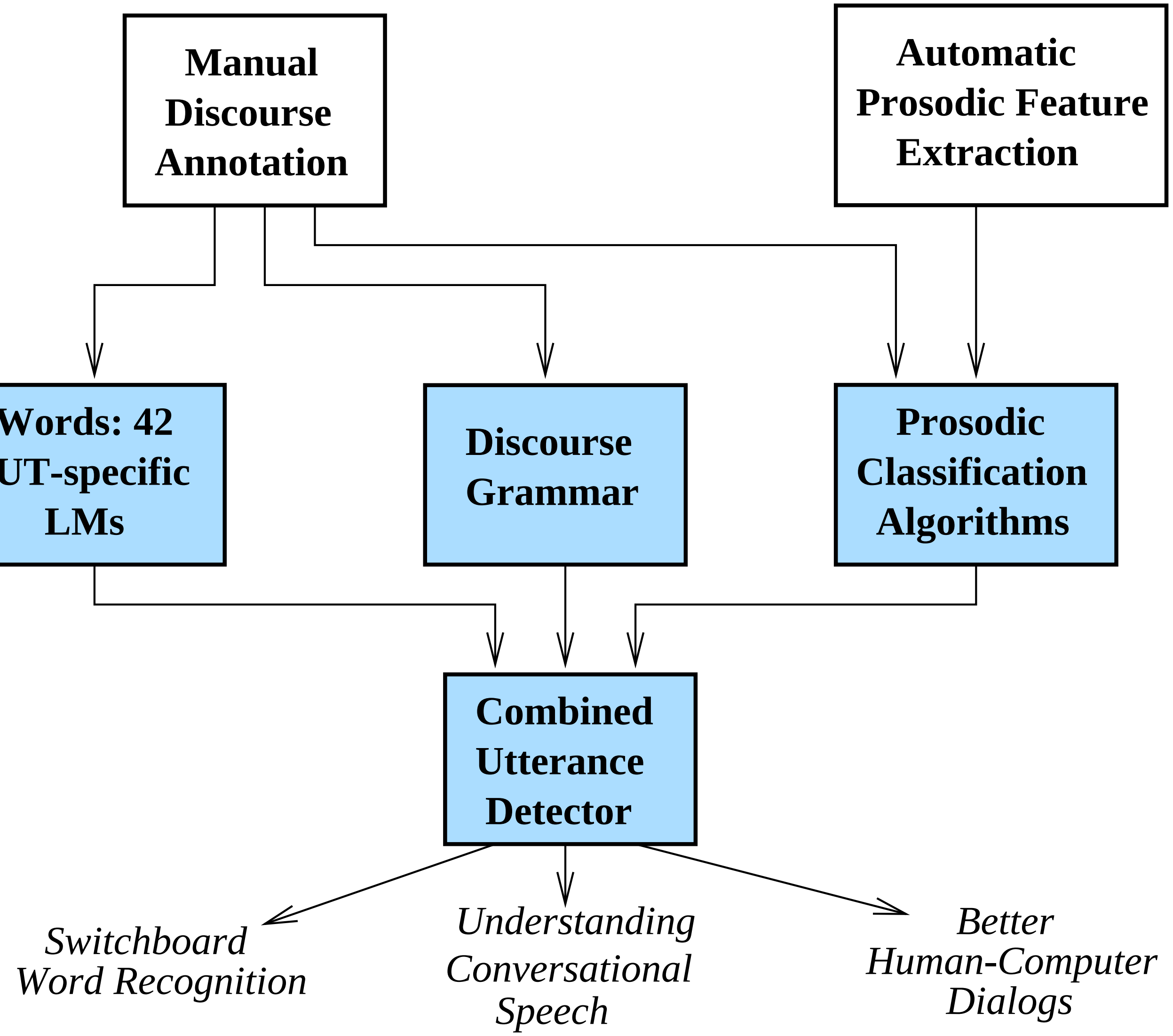
Examples: Abandoned and Turn Exits

A	**B**	A	*B*
Statement		we’re from, uh, I’m from Ohio /	
Statement		and my wife’s from Florida /	
Turn-Exit		so, -/	
	Backchannel		*Uh-huh. /*
. . .			
Hedge		so, I don’t know, /	
Abandoned		it’s <lipsmack>, - /	
Statement		I’m glad it’s not the kind of problem I have to come up with an answer to because it’s not –	

MOVING RIGHT ALONG

Part II: Utterance Type Detection

Where are we?


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

Reminder: Why Utterance Type Detection?

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 just got asked a question or ordered to do something.

Better Switchboard Word Error:

Different Utterance types have very different words.

Reminder: Hand Segmented Test Set

- This summer no focus on automatic segmentation.
- But not ignored:

SRI: Stolcke and Shriberg 1996 (85% recall, 70% precision given correct words)

CMU: Gavalda et al 1996, Buo and Waibel 1996

Others: Grosz and Hirschberg 1992, Hearst 1994,1997, Hirschberg and Nakatani 1996, Ostendorf and Veilleux 1994, Passonneau and Litman 1993,1995, many others

Algorithms: Decision trees, ANN, decision lists, feature induction, etc.

How well do people do at this task?

- Human agreement **83%**
- **Kappa Statistic** *(Carletta 1996, Flammia and Zue 1995, Isard and Carletta 1995)*

$$\kappa = \frac{\text{Percent\_Agree} - \text{Percent\_Expected}}{1 - \text{Percent\_Expected}}$$

- Human labelers pairwise kappa is **.80**.
- Pairwise scores .7 or higher = good reliability
  Pairwise scores .8 or higher = high reliability

Utterance Type Detection

Words and Word Grammar: Pick the Utterance Type which is most likely given the word string *(Garner et al 1996, Hirschberg and Litman 1993, etc)*

Discourse Grammar: Pick the Utterance Type which is most likely given the surrounding UTs. *(Kita et al 1996, Reithinger and Maier 1995, Reithinger et al 1996, Sarukkai and Ballard 1996, Suhm and Waibel 1994, Woszczyna and Waibel 1994, Yamaoka, and Iida 1990, Taylor et al 1996, 1997*

Prosodic Information: Use pitch contour, energy, SNR, speaking rate to choose Utterance Type *(cf. Takagi and Itahashi 1994, Terry et al 1994, Waibel 1988, Taylor et al 1996, 1997)*

Word-based Utterance Type Detection

Utterances are distinguished by their words and word sequences

- **92.4%** of the **"uh_huh"**'s occur in **Backchannels**
- **50.2%** of the **"do you"**'s occur in **Yes-No-Questions**
- **88.4%** of the trigram "<**s**> **do you**" occur in **Yes-No-Questions**

Prosodic-based Utterance Type Detection

- Need to distinguish **Questions** from **Statements**
- Words aren't that different in reference strings.
- Don't have reference strings!
- Initial words in lattice often deleted!
- So word-based detector gets only **32%** accuracy
- Yes-No-Questions have final F0 rise.

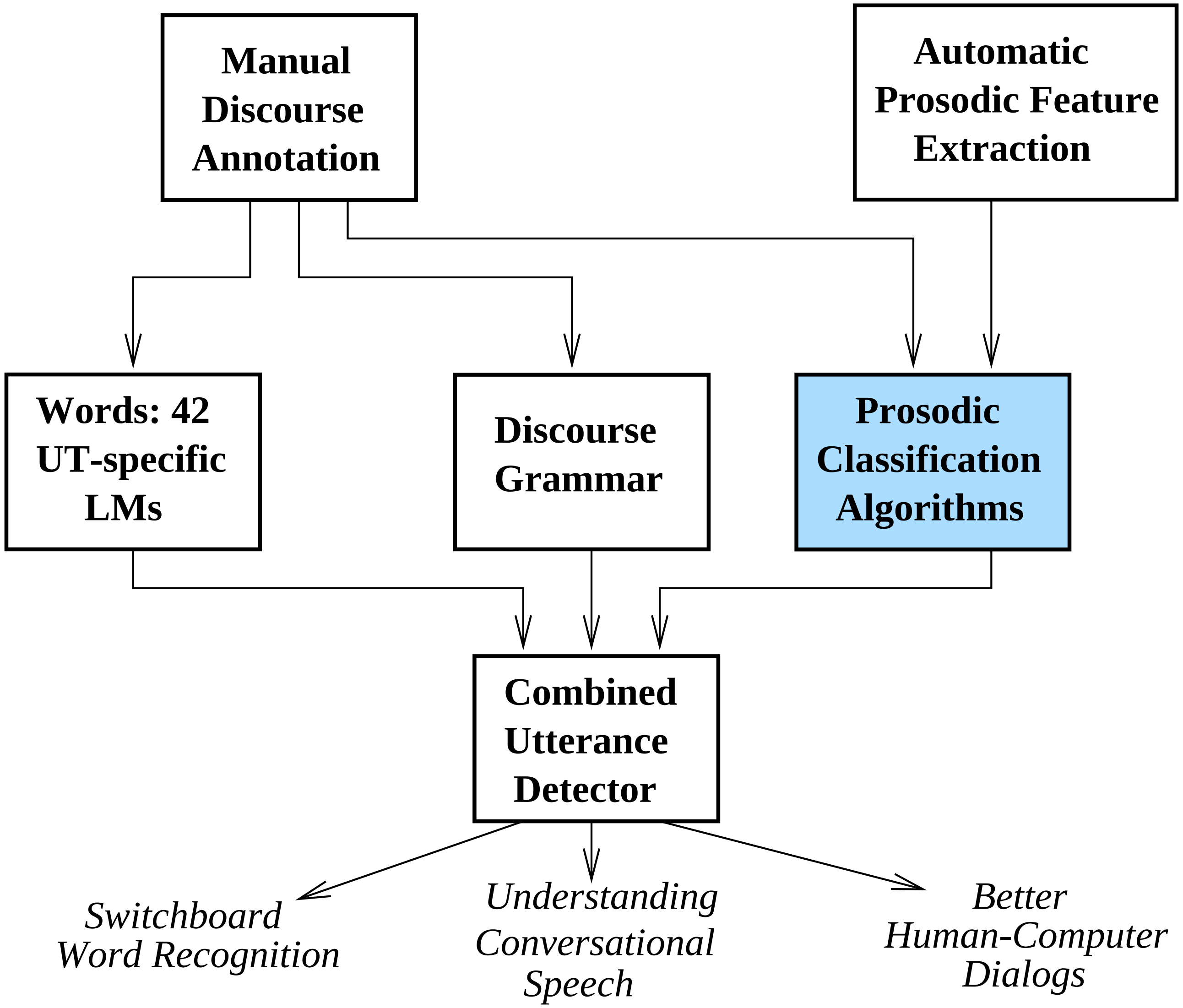
Confusable Utterance types: Can Prosody help?

- LM has trouble distinguishing **Backchannel**, **Answer-Yes**, and **Agree**
- Because all are often *yeah* or *uh-huh*.
- Important to distinguish since help disambiguate neighboring UTs
- For example **Answer-Yes** is a terrific cue to a prior question
- Can use **Stress** to help distinguish

Discourse Grammar Utterance Type Detection

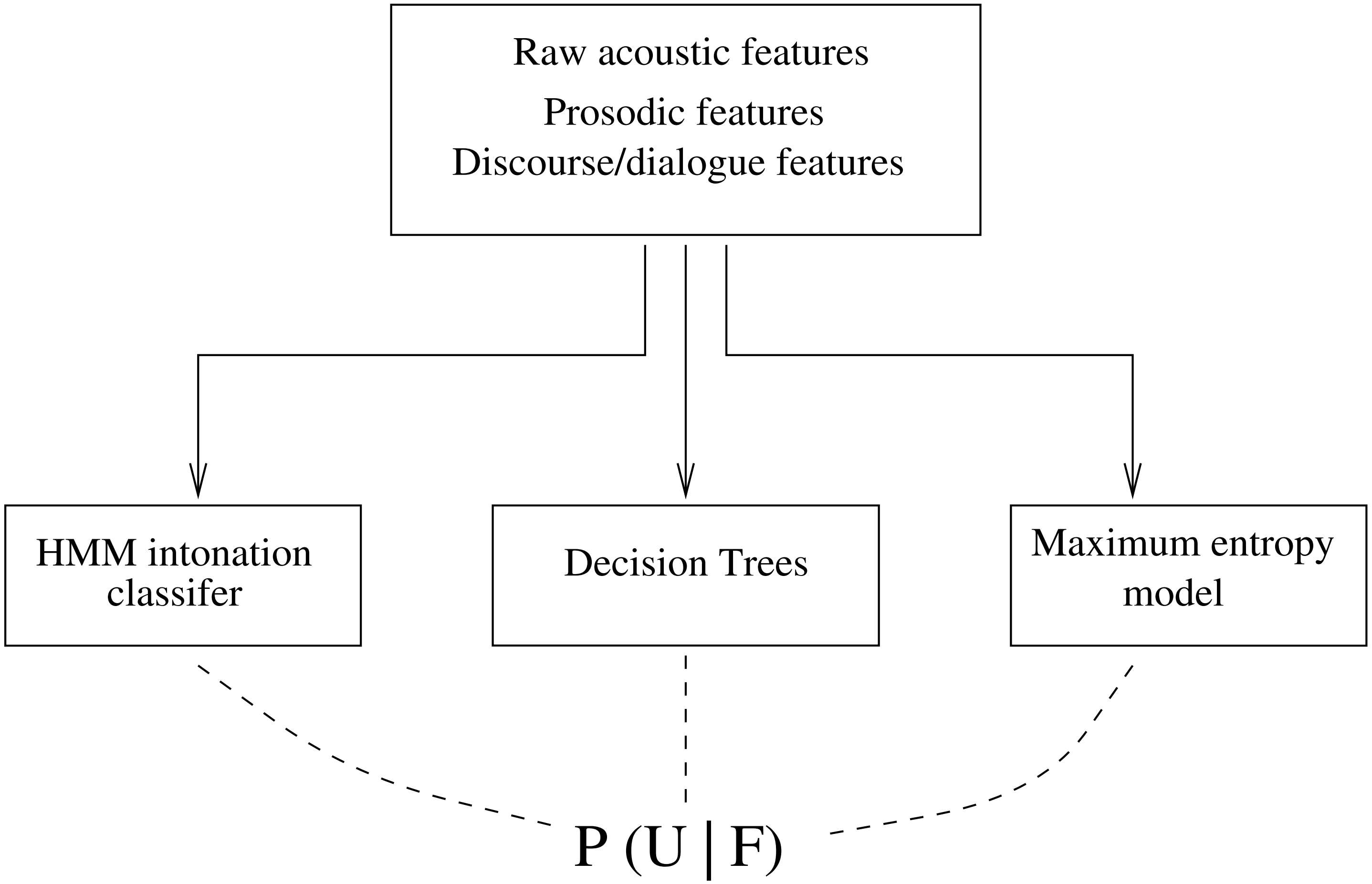
- Bigram probability that a **Yes-No-Question** will receive a **Yes** answer: **.30**
- Bigram probability that a **Yes-No-Question** will receive a **No** answer: **.14**
- Bigram probability that a **Command** will be **Agreed** to: **.23**
- Bigram probability that a **Statement** will be **Backchanneled**: **.26**

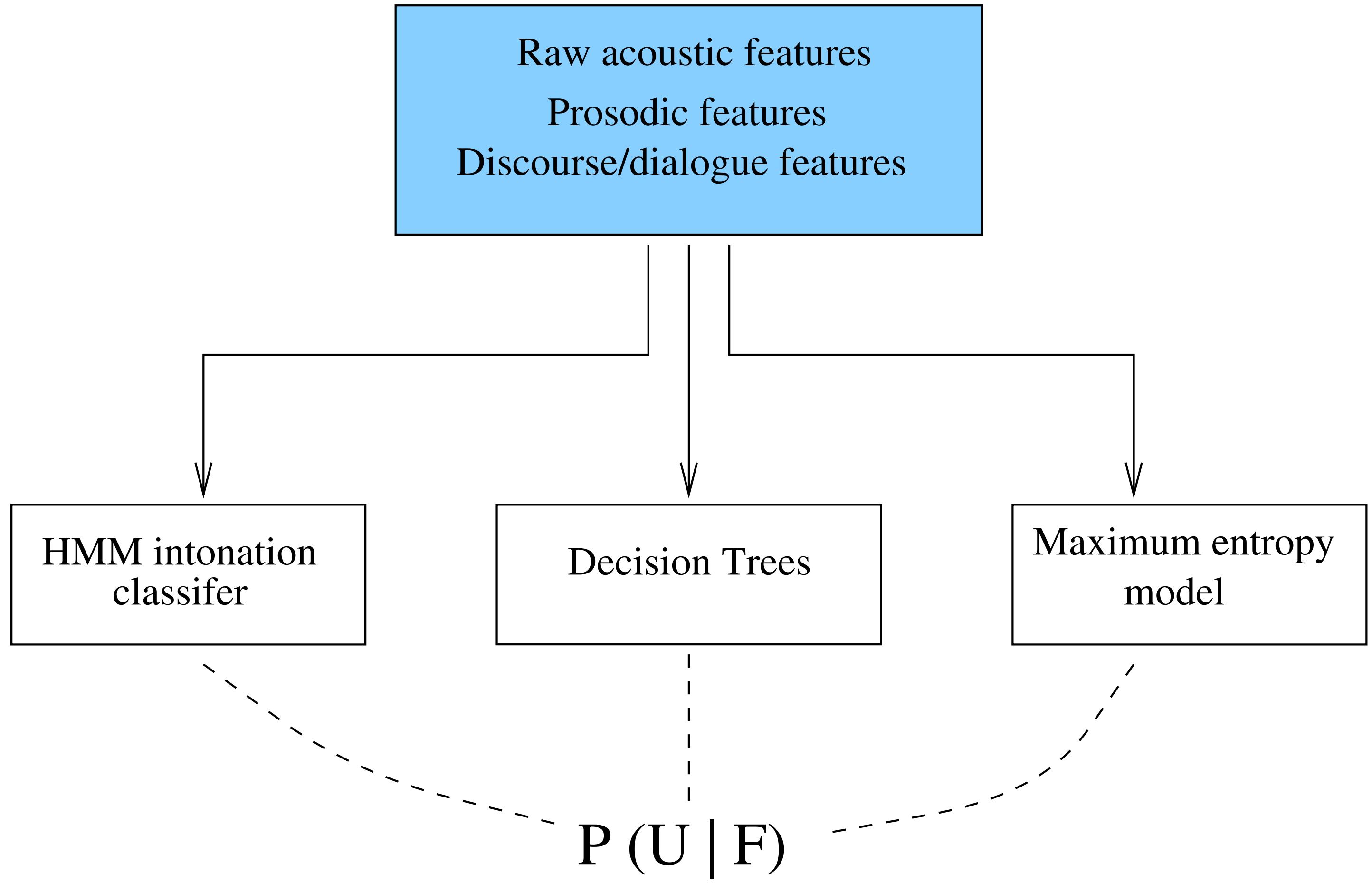
Where are we?



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

Utterance Type Detection from Acoustic and Prosodic Information

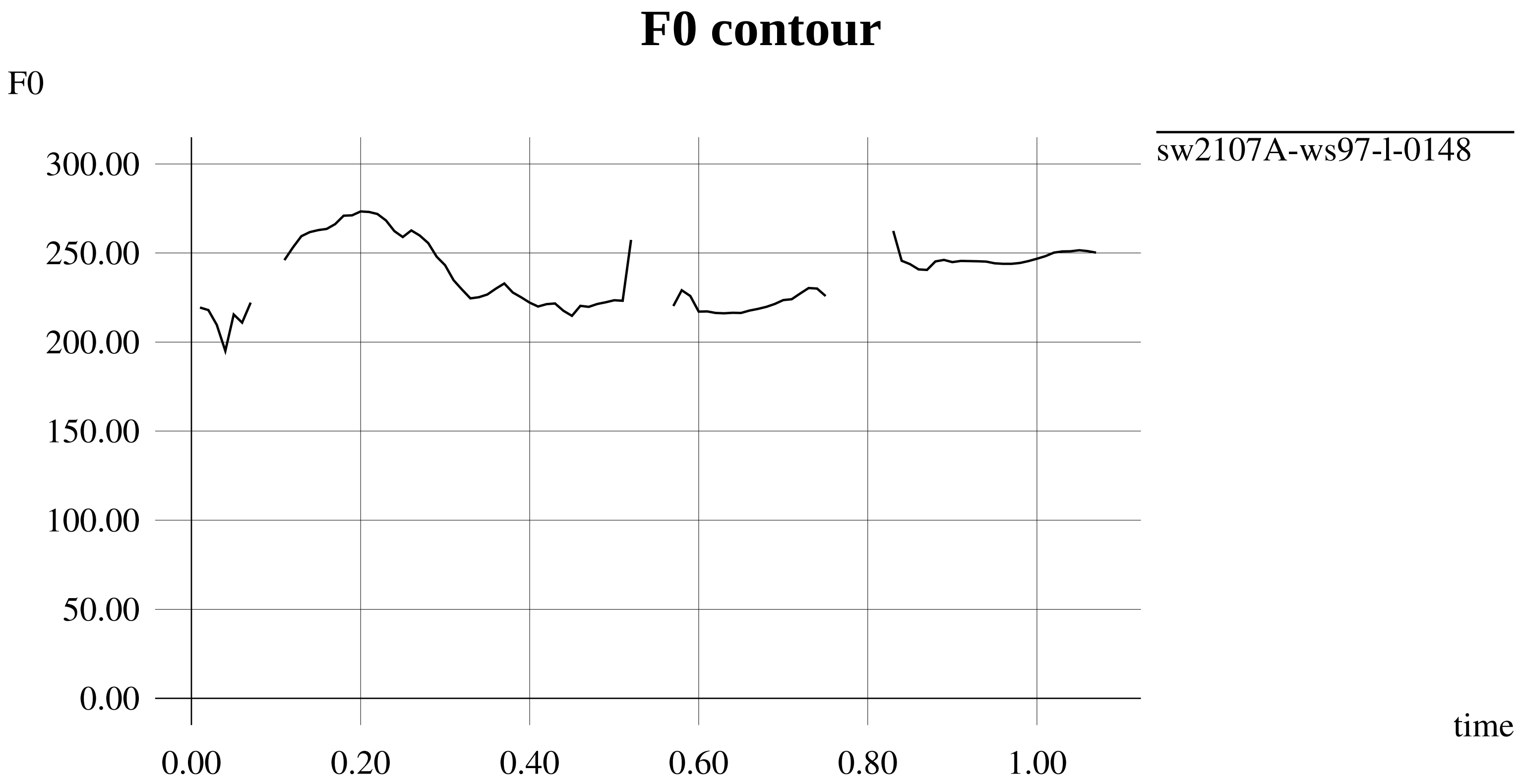



Raw acoustic features
Prosodic features
Discourse/dialogue features
HMM intonation classifer
Decision Trees
Maximum entropy model
P (U | F)

Features used

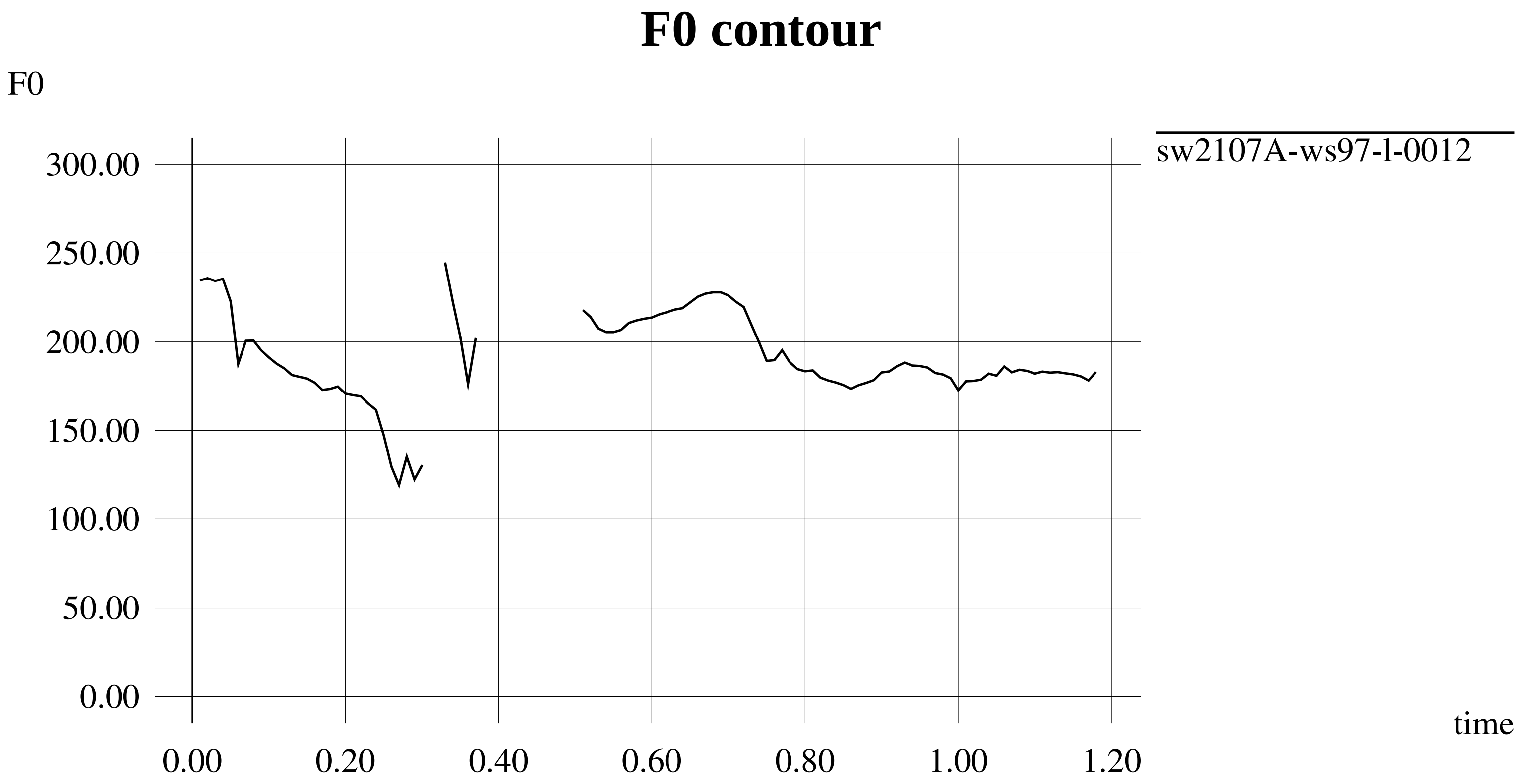
- Raw Acoustic Features
  - RMS energy
  - Estimation of signal-to-noise ratio
  - F0 contour, both raw and median smoothed
  - Estimation of speaking rate (enrate)
  - Utterance duration
- Derived Acoustic Features
  - Normalised, max, min, gradient and mean values
  - Special analysis of end sections of utterance, energy and F0
  - pitch accent detector (simple HMM system trained on F0 and energy)
- Features associated with discourse/dialogue
  - Turn initial?
  - Number of interrupts (on other channel)
  - Gender

F0 contour for yes/no question



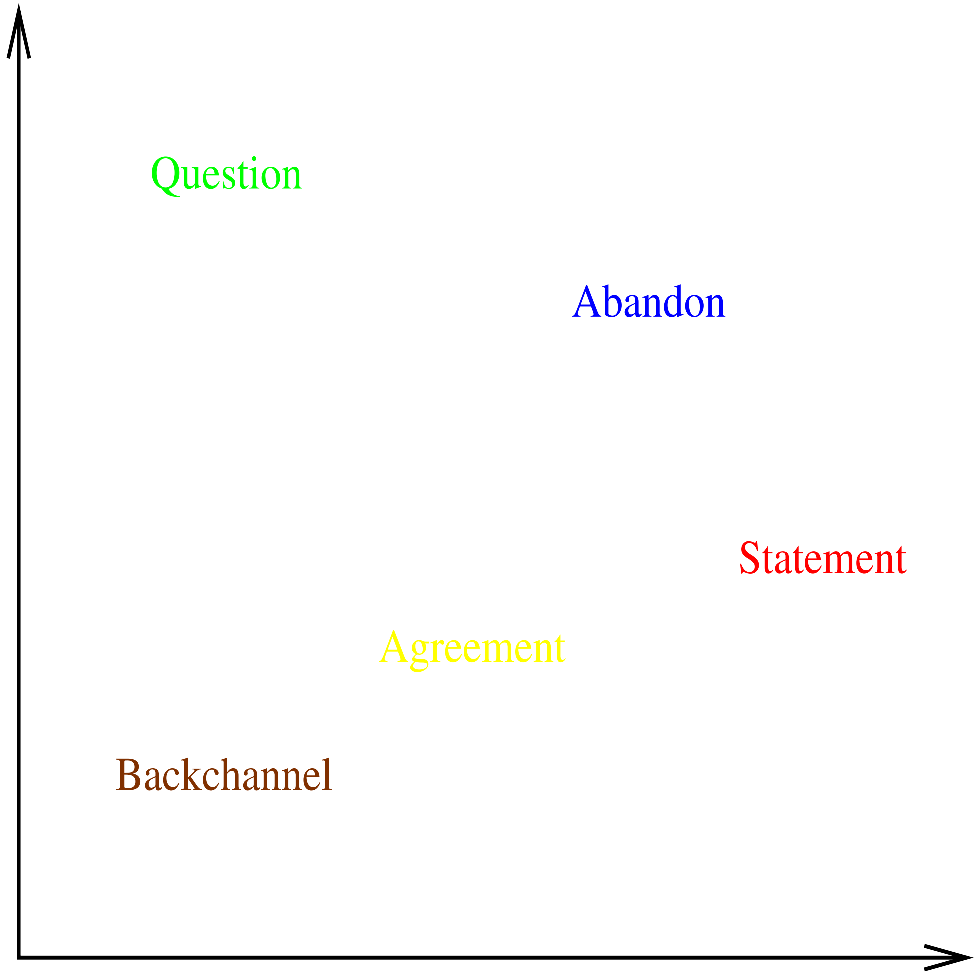
Do you ever read cosmo?

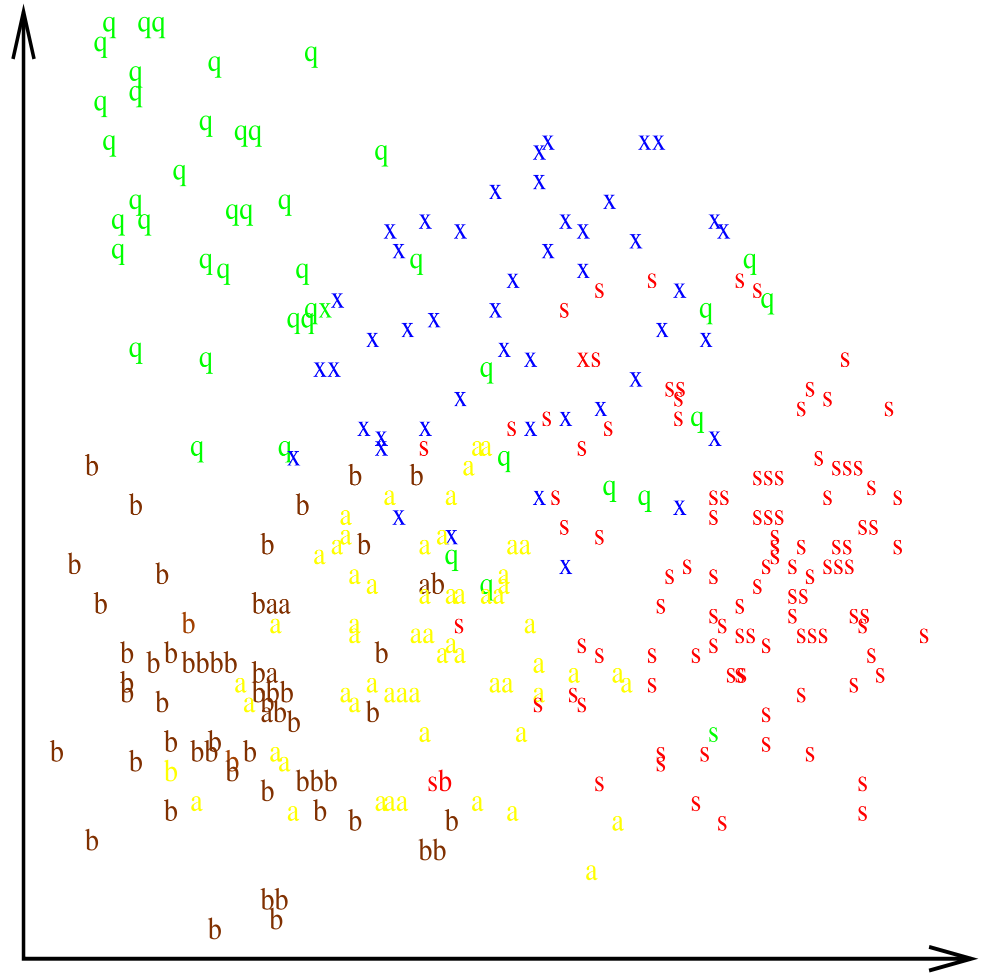
F0 contour for statement



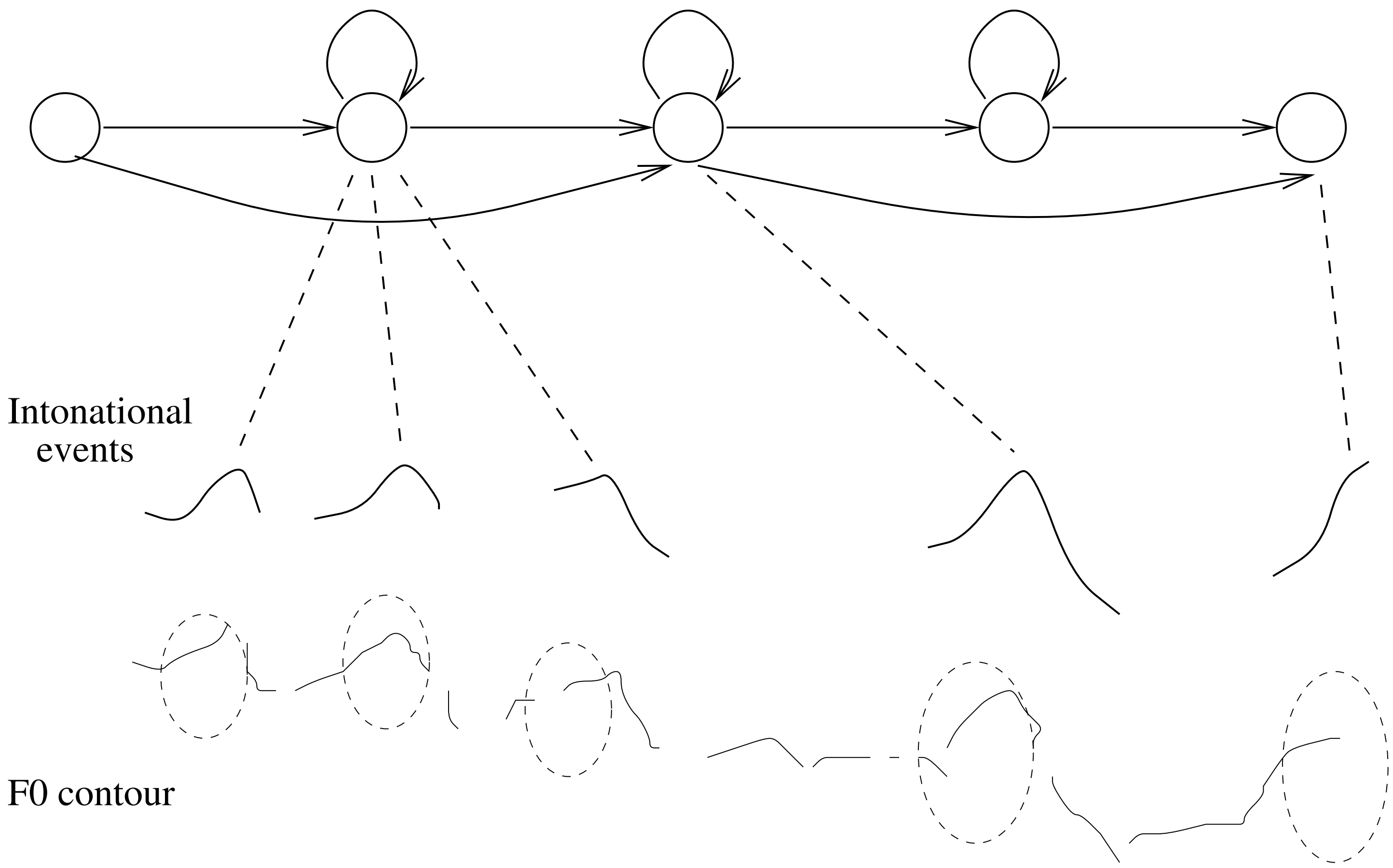
It started raining.

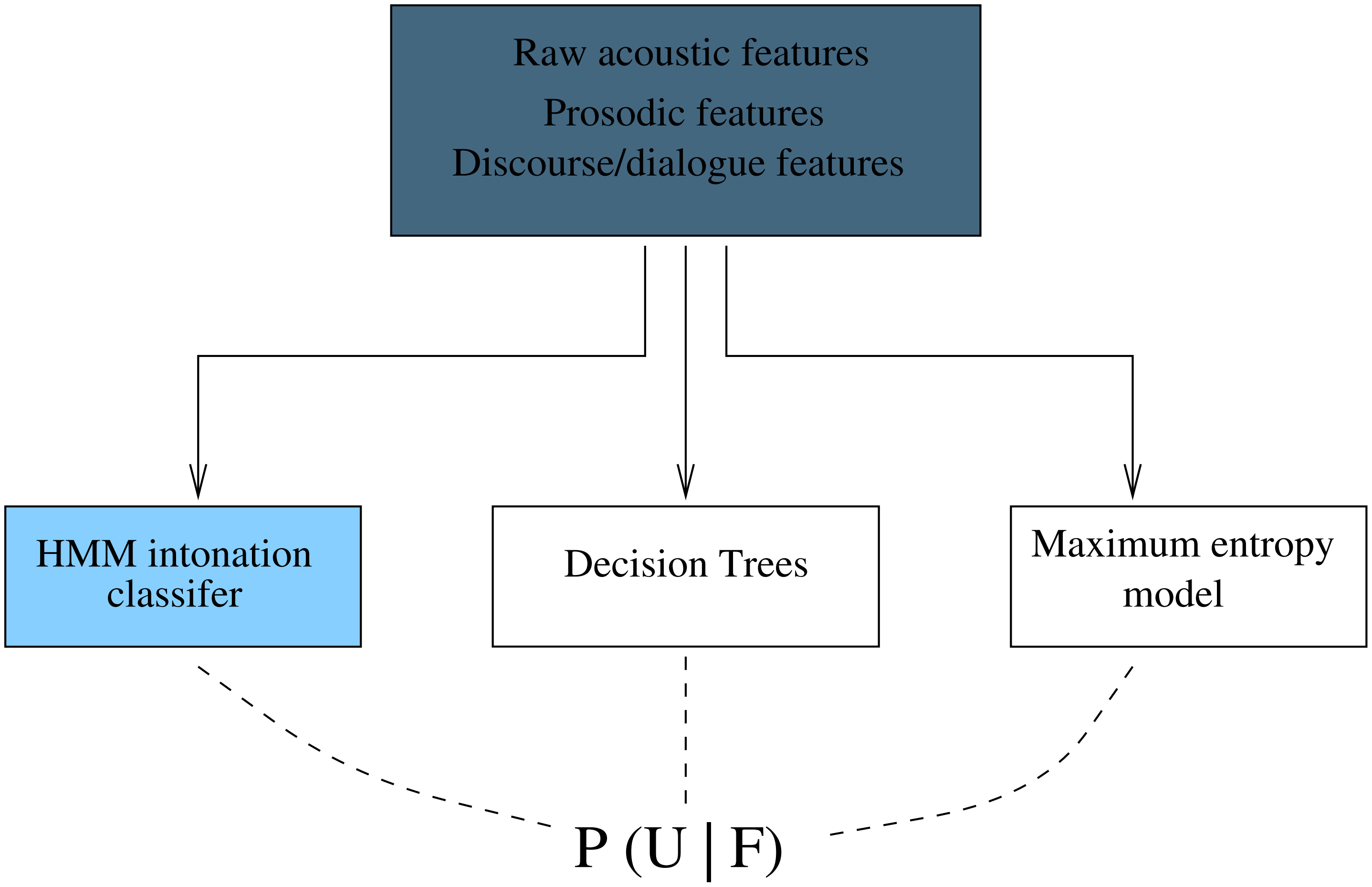
Distributions of 5 Main Utterance types





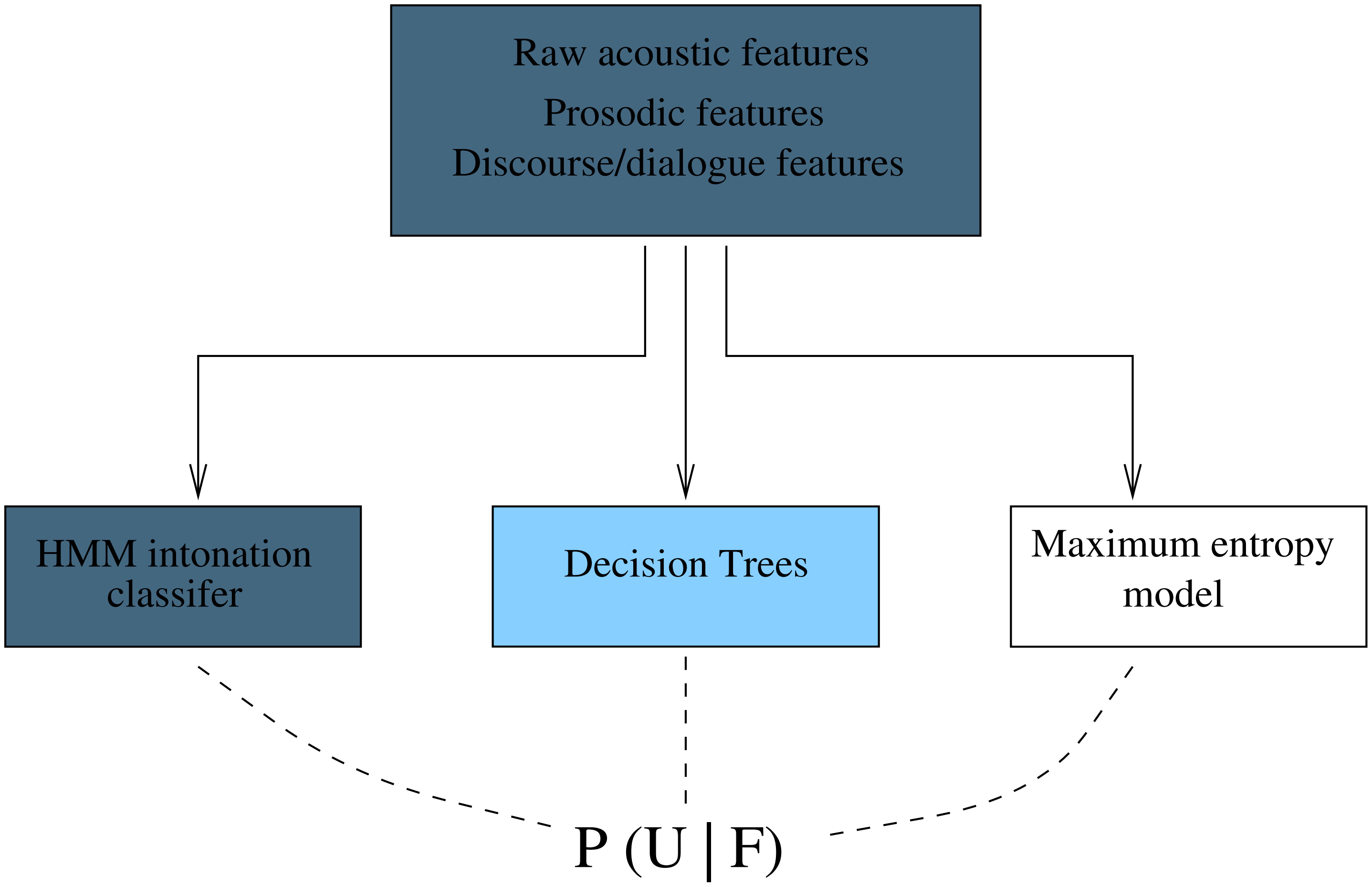
HMM model




Raw acoustic features
Prosodic features
Discourse/dialogue features
HMM intonation classifer
Decision Trees
Maximum entropy model
P (U | F)

HMM Classification

- We know that it is often the *change* of a particular feature throughout the utterance which signals distinctive behaviour.
- Approach: Train one HMM per utterance type, the HMMs being trained on all the observations for the utterance.
- System 1: Use sequences of parameterized pitch accents as observations. Accuracy **41%.**
- System 2: Use sequences of F0 and RMS energy values (+deltas) as observations. Accuracy **45%.**


Raw acoustic features
Prosodic features
Discourse/dialogue features
HMM intonation classifer
Decision Trees
Maximum entropy model
P (U | F)

Decision Trees

Goals:

- discriminate confusable classes
- understand feature usage and combination

Can we detect questions?

Motivations:

- Following utterance depends on distinction.

  We get donuts again? **Yeah.**

  We get donuts again. **Again?!**

- Confusable based on recognized words:

	Classification Rate
All Types	63.5%
Questions	32%

Method

- IND, CART-style tree
- downsampled to equal class priors (50% Q, 50% Other)
- can adjust for priors in integration phase
- DLM train set = 1788 sides
- 10-fold cross-validation $\rightarrow$ pruned tree
- apply tree to 2 test sets
  - WS97DEV 38 sides
  - held out 436 sides
- accuracy = no. correct classifications / all data
- downsampling means chance is 1/N or 50%

Detecting Questions: Results?

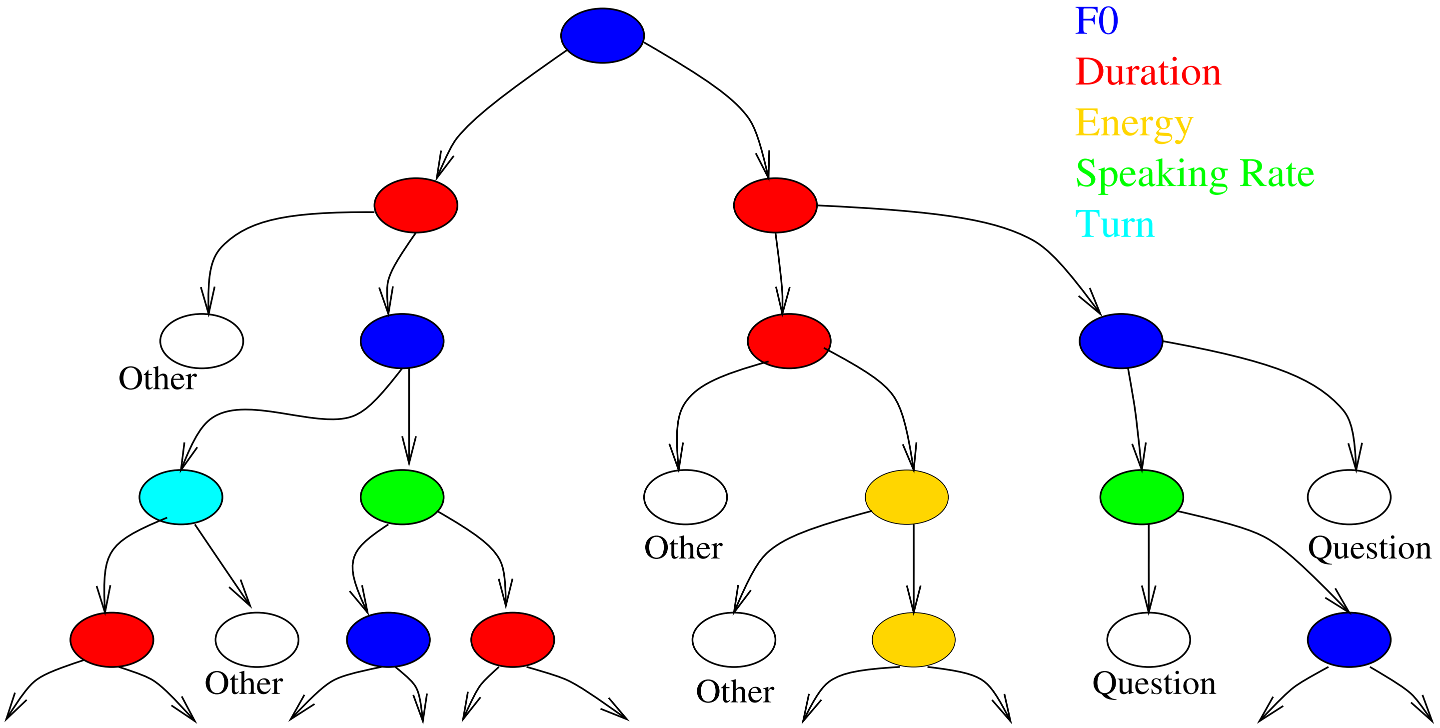
Test	Count	Accuracy	Perplexity	Entropy	Efficiency
Prior		50.0	2	0.693	
Cond. HLD	618	70.3	1.80	0.589	15.0%

Not enough cases in devtest!

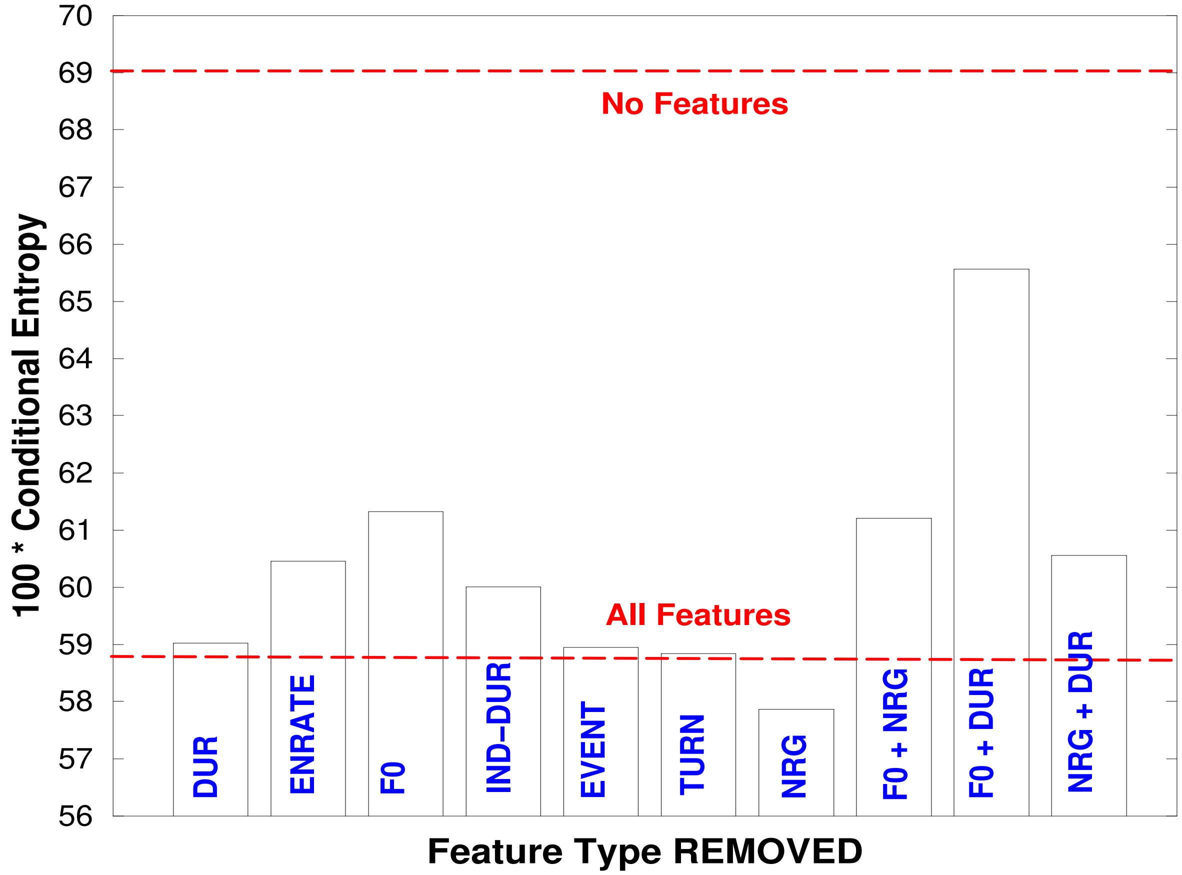
Which features do the work? Groups:

- F0
- duration
- energy (raw, SNR)
- speaking rate (enrate)
- event (accents, boundary tones)

Question Detection



Feature Importance



Detecting Abandoned Utterances

So, then he –

Based on recognized words:

- average classification 63.5%
- for abandonments: $\approx$ 21%

Test	Count	Accuracy	Perplexity	Entropy	Efficiency
Prior		50.0	2	0.693	
Cond. HLD	1363	75.5	1.67	0.514	25.8%
Cond. Dev	183	70.5	1.80	0.590	14.9%

Better performance for HLD could be due to:

- Higher variability in small sample
- mismatch between train and devtest

Detecting of Abandonments

Is good performance just due to length?

Removed utt types with longest counts
(i.e., full statements and full questions)

Test	S&Q?	Count	Acc.	Perp.	Entropy	Efficiency
Prior			50.0	2	0.693	
Cond. HLD	YES	1363	75.5	1.67	0.514	25.8%
Cond. Dev	NO	1340	77.6	1.69	0.526	24.1%

Uses variety of features.

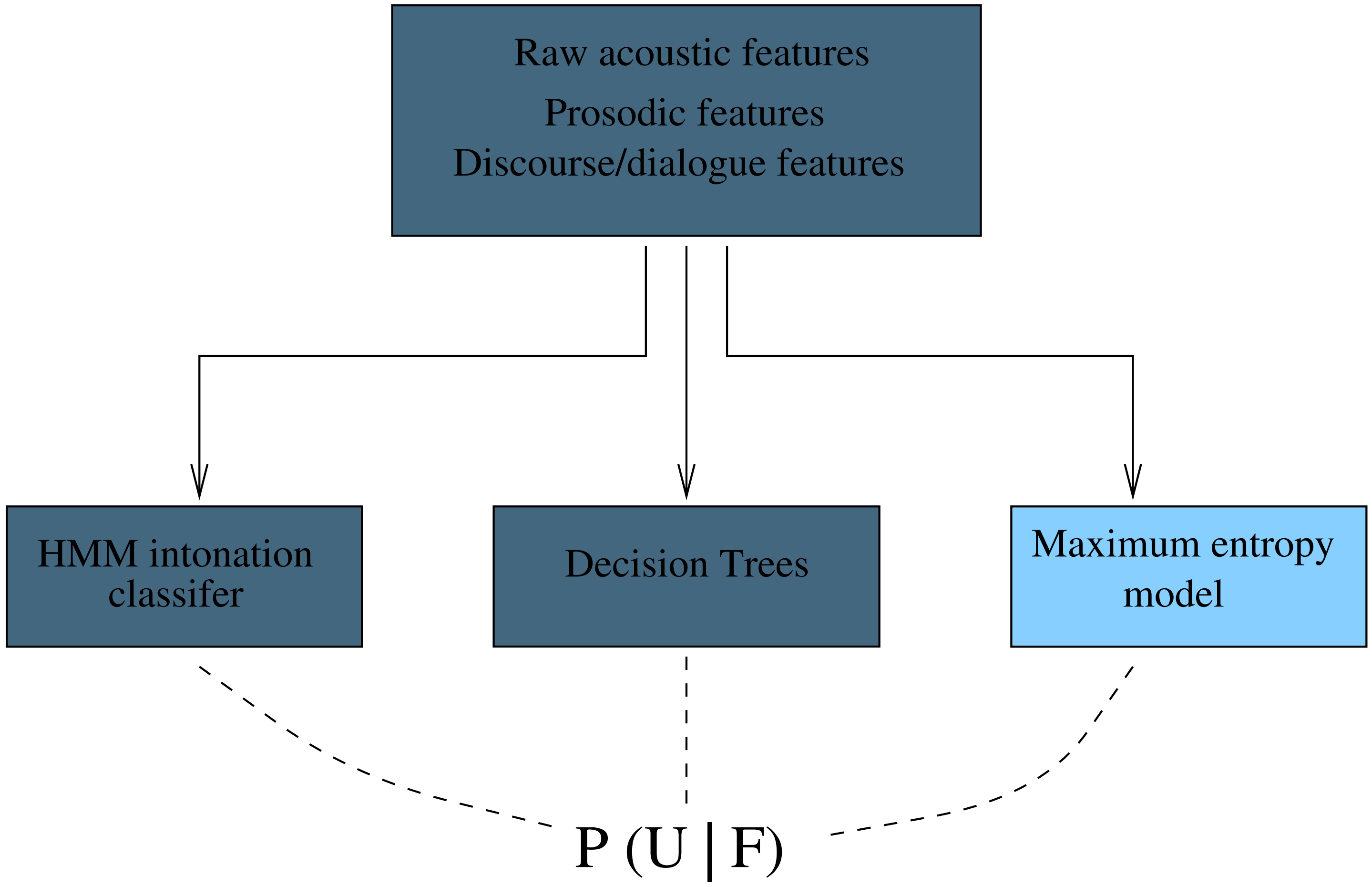
Example: F0 regression over utterance

- Less steep falls $\rightarrow$ abandoned
- Steep falls $\rightarrow$ nonabandoned

Classification of 5 Most-Frequent Utterance Types

Statements, Questions, Backchannels, Agreements, Abandoned, Other

Test	Count	Accuracy	Perplexity	Entropy	Efficiency
Prior		16.7	6	1.79	
Cond. HLD	1269	45.1	3.85	1.34	24.8%
Cond. Dev	107	37.7	4.25	1.45	19.2%


Raw acoustic features
Prosodic features
Discourse/dialogue features
HMM intonation classifer
Decision Trees
Maximum entropy model
P (U | F)

Exponential Models vs. Decision Trees

Two types of explanations

- decision trees are a computational model
- humans can interpret (small) decision trees
- mathematically they might **NOT** explain the underlying feature interactions

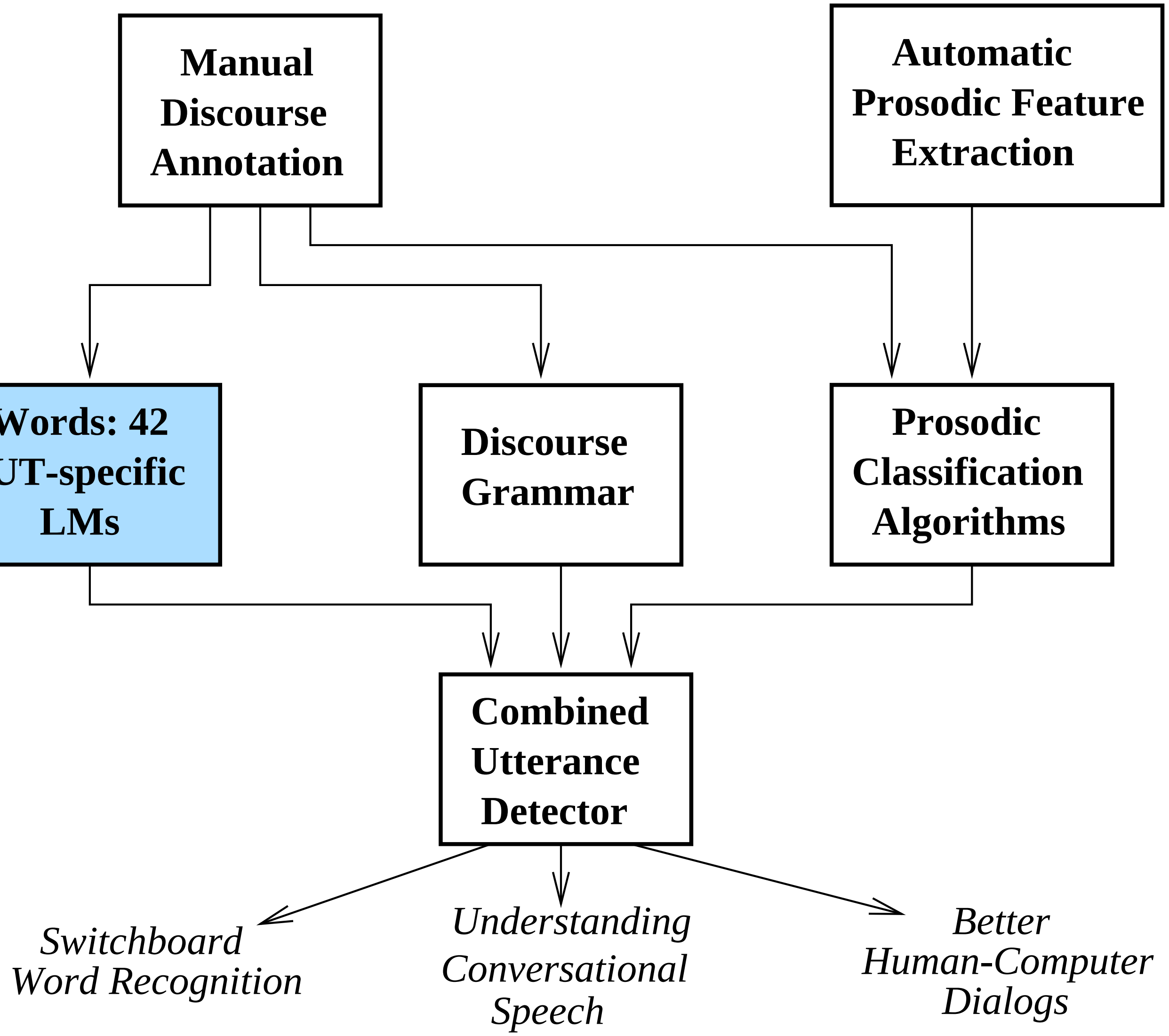
$\rightsquigarrow$ try something else, e.g. exponential models:

- discrete features same as for Decision Tree
- continuous features are made discrete by binning
- no induction of composite features

Prosodic Features: Conclusions

- prosodic features help distinguish otherwise confusable classes
- decision trees use a variety of features
- F0 and duration are very important
- consistent discrepancies between HLD and devtest
  - need good utterance boundary times for training data
  - this should improve results
- good features but need to explore other classifiers

Where are we?


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

Modeling Discourse

language model	*discourse grammar*
But I turned out to be a White Sox fan.	Statement
Well, that's all right.	Downplayer
I lived in Chicago.	Statement
How do you expect them to do this year?	Open-Question
Yeah.	Backchannel
They did pretty well.	Statement
Okay.	Response-Acknowledge
What did you think of the World Series?	Open-Question

42 Utterance Type-Specific Language Models

- one trigram model per utterance type (UT) $\leadsto$ 42 models
- linear interpolated with trigram built on all of SWBD

LM	Perplexity
Baseline	76.8
UT-specific	66.8

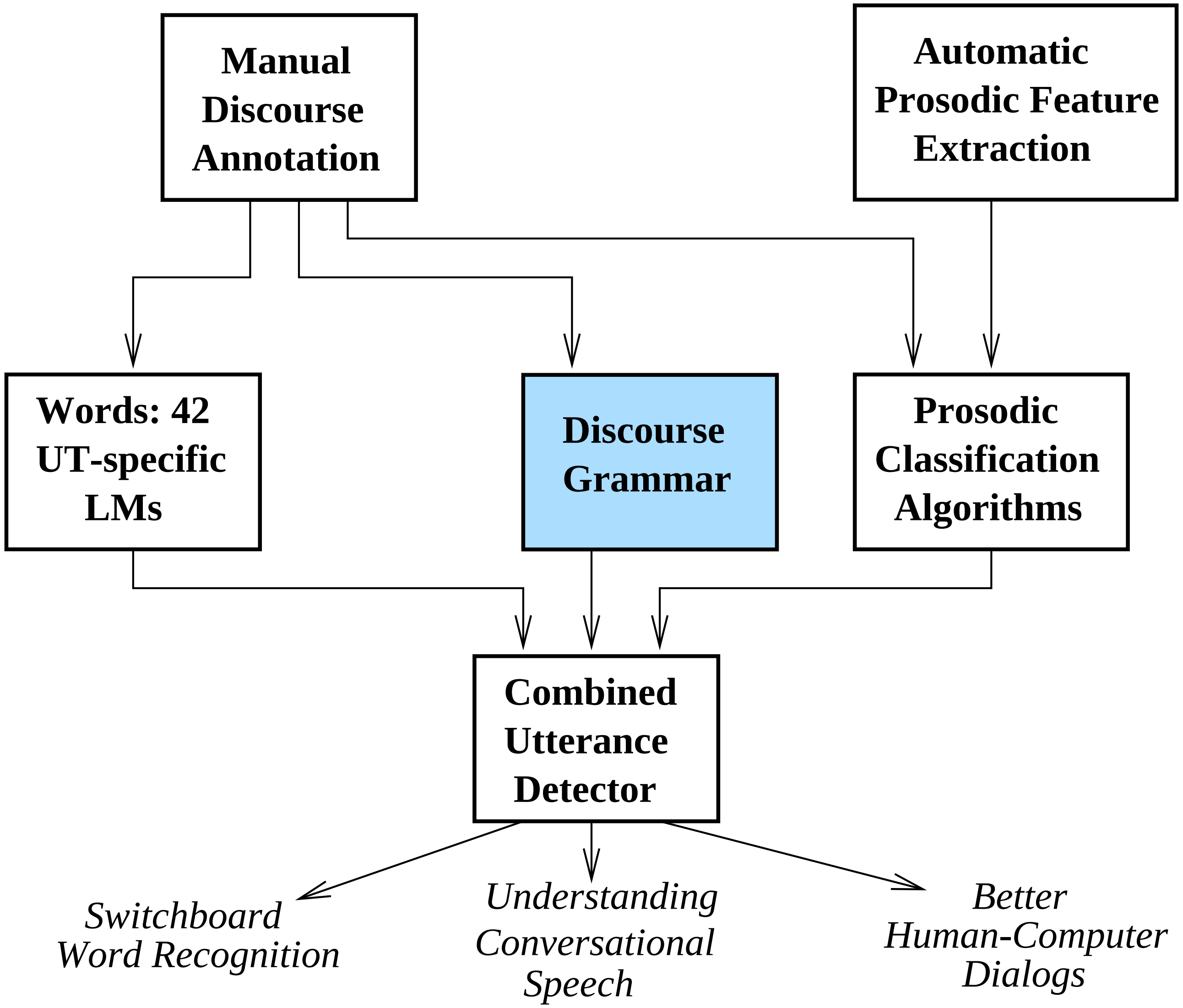
Utterance Type Detection

Choose most likely UT given the word string, uniform utterance prior.

$$P(\text{ I lived in Chicago.} \mid \texttt{Statement})$$

UT Prediction Accuracy		
chance	human transcripts	recognizer output
35%	54.3%	42.8%

Where are we?



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

Discourse grammar

Use ngram models to model sequences of utterance type/speaker pairs:

Channel A `Statement` `Statement` `Backchannel` `Statement`

Channel B `Downplayer` `Open-Question`

	n-gram model			
n	0	1	2	3
perplexity	42.0	9.8	5.5	5.2

A Random Conversation

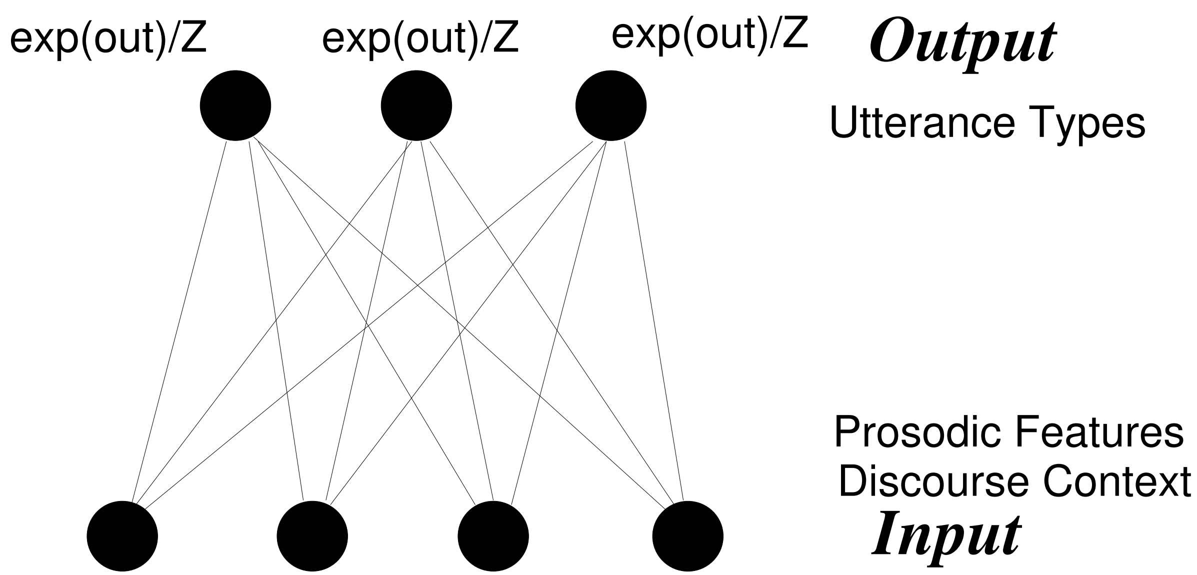
Backchannel	Uh-huh.
Statement	And, it was so good.
Backchannel-Q	Really.
Statement	But when we found it, it was, we knew it.
Statement	and as luck had it, uh, she had puppies the same week my daughter was born, /
Backchannel	Uh-huh. /
Statement	I hear that from, my brother-in-law lives in Plano,
Backchannel	Right,
Statement	And I like Demi Moore.
Statement	I just still have need for the four bedrooms just having company from time to time.
Appreciation	you're not kidding.
Agree	Absolutely.
Agree	Absolutely.

Discourse Grammars are different

- small size of vocabulary – *42 types*
- small training set – *one tag per utterance*
- overlapping features are interesting
- candidate models are
  - Exponential Models (Maximum Entropy)
  - Neural Networks
  - Decision Trees

Exponential Models (Maximum Entropy)

- n-grams & generalizations
- results are identical with regular trigram
- could integrate prosodic features directly

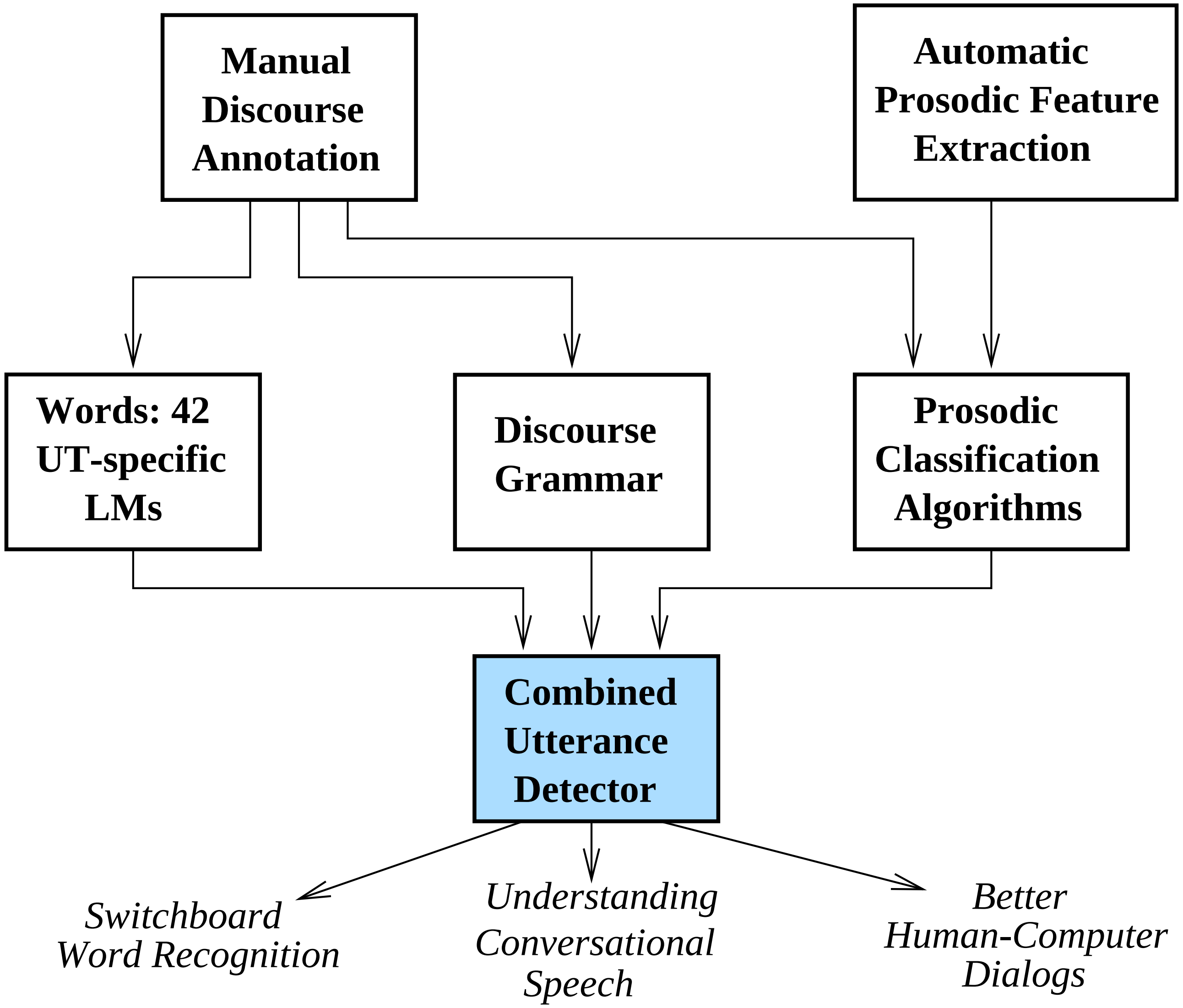


Lambda values for the LM-interpolation

Utterance Type	LAMBDA
Backchannel	0.981439
Response-Acknowledge	0.978348
Answer-No	0.969542
Answer-Yes	0.946166
Backchannel-Question	0.93424
Appreciation	0.868793
Agree	0.826063
Abandoned	0.815712
Thanks	0.750884
...	...
Answer Other	0.256439
Quotations	0.245256
Summarization/Reformulation	0.226074
Answer Maybe	0.225677
Declarative Wh-Questions	0.18137
Disprefered Response	0.178244

Part III: Putting it all Together

Where are we?



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

Model Integration I: UT Detection from Words

Idea: Combine local evidence (UT likelihoods) with a discourse LM to find globally best UT sequence U.

Simplest Case: Word sequence W is given

$$U^* = \operatorname*{argmax}_U P(U|W) = \operatorname*{argmax}_U P(U)P(W|U)$$

$P(U)$ is estimated using a N-UT-gram discourse grammar

Assume words independent across utterances, given UTs:

$$P(W|U) = P(W_1|U_1)P(W_2|U_2)\dots P(W_n|U_n)$$

Each $P(W_i|U_i)$ is given by UT-specific trigram LM.

UT Detection from Words: Results

UT detection accuracies on 19 devtest conversations, using full conversation transcripts:

Discourse Grammar used	Accuracy (%)
None	53.9
Unigram	68.9
Bigram	71.4
Trigram	71.8

Note: UT-specific LMs were *not* interpolated with the general LM to increase discrimination.

Model Integration II: UT Detection using Recognizer

Idea: Local UT likelihoods now based on acoustics A.

$$\begin{aligned} U^* &= \operatorname*{argmax}_U P(U|A) \\ &= \operatorname*{argmax}_U P(U)P(A|U) \\ &= \operatorname*{argmax}_U P(U) \sum_W P(A|W)^{1/\lambda} P(W|U) \end{aligned}$$

$P(A|W)^{1/\lambda}$ is recognizer acoustic score (UT independent), scaled down to match dynamic range of LM ($\lambda =$ language model weight).
$P(W|U)$ is UT-specific LM.
Summation over all W approximated by N-best list.

UT Detection using Recognizer: Results

UT detection accuracies on 19-conversation devtest set, scored utterances only. Bigram discourse grammar.

Evidence used	Accuracy (%)
1-best recognition	63.5
2500-best recognition	64.6
Reference transcripts	71.0

Note: UT detection from reference transcripts is harder than from full transcripts due to missing utterances and truncation after 5 minutes.

Model Integration III: UT Detection from Prosody

Idea: Local UT likelihoods now based on prosodic features F.

$$\begin{aligned} U^* &= \operatorname*{argmax}_U P(U|F) \\ &= \operatorname*{argmax}_U P(U)P(F|U) \end{aligned}$$

Assume prosody independent across utterances, given UTs:

$$P(F|U) = P(F_1|U_1)P(F_2|U_2)\dots P(F_n|U_n)$$

Each $P(F_i|U_i)$ is given by a decision tree trained using uniform priors:

$$P(F_i|U_i) = P(F_i)\frac{P(U_i|F_i)}{P(U_i)} \propto P(U_i|F_i)$$

UT Detection using Prosody: Results

UT detection accuracies on 19-conversation devtest set, scored utterances only.

Note: Very preliminary results using one tree that distinguishes 5 most frequent UTs and “other” classes.

Discourse Grammar	Accuracy (%)
None	38.9
Unigram	48.3
Bigram	50.2

Model Integration IV: Using Recognizer and Prosody

Idea: Local UT likelihoods based on both acoustics and prosodic features F, assuming independence given U.

$$
\begin{aligned}
U^* &= \operatorname*{argmax}_U P(U|A,F) \\
&= \operatorname*{argmax}_U P(U)P(A,F|U) \\
&\approx \operatorname*{argmax}_U P(U)P(A|U)P(F|U)
\end{aligned}
$$

$P(A|U)$ approximated as before using N-best hypotheses.

$P(F|U)$ estimated as before using decision tree.

UT Detection using Recognizer and Prosody: Results

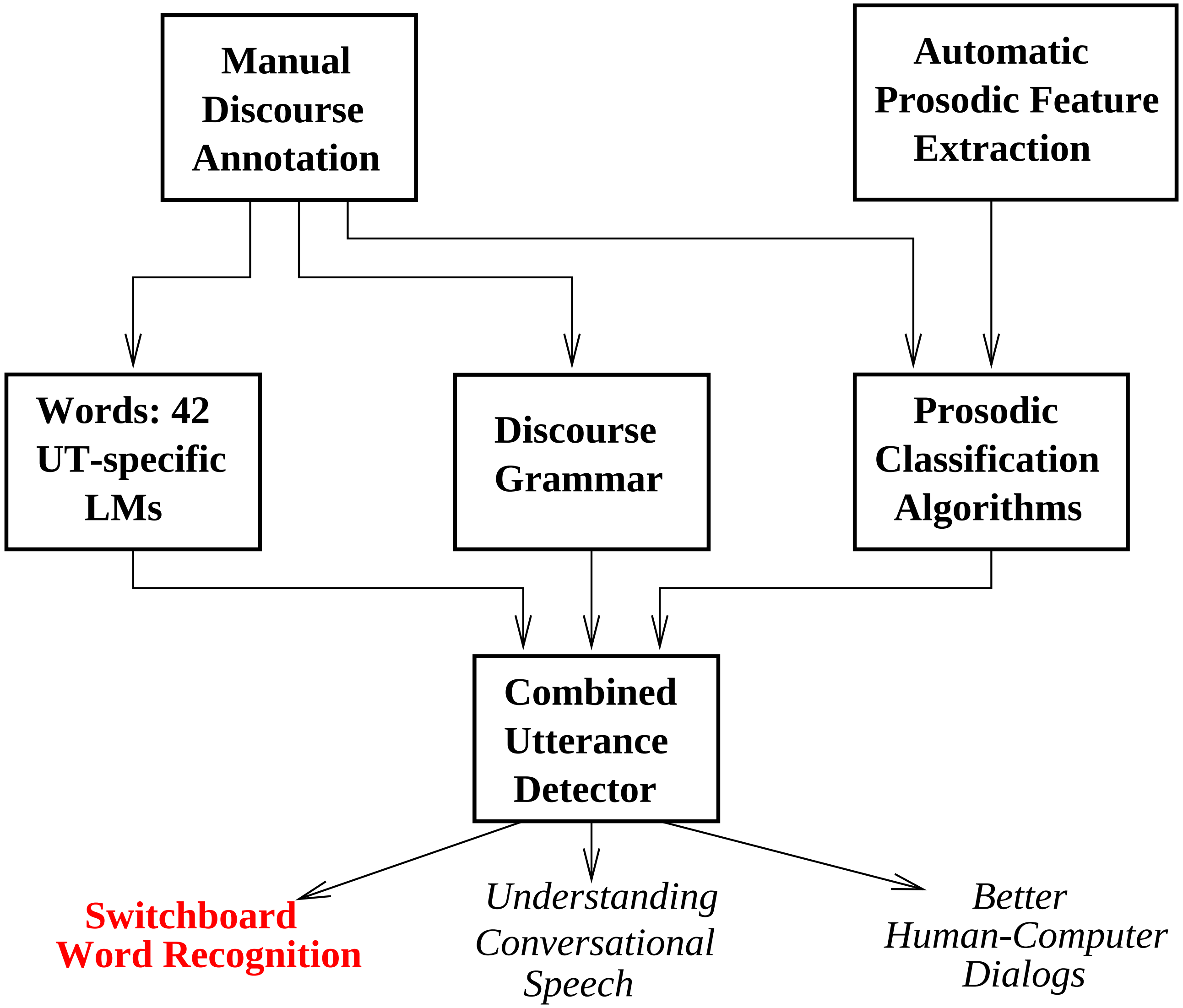
UT detection accuracies on 19-conversation devtest set, scored utterances only.

	Accuracy (%)		
Discourse Grammar	Prosody only	Recog. only	Combined
None	38.9	42.8	56.5
Unigram	48.3	61.9	62.6
Bigram	50.2	64.6	65.0

Why are combined likelihoods not helping with discourse grammars?

Are independence assumptions grossly violated?

Where are we?



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

Recognition with Discourse Grammar

Recognize with a weighted mixture of UT-specific language models.
Weights are given by posterior UT probabilities.

$$
\begin{aligned}
W_i^* &= \operatorname*{argmax}_{W_i} P(W_i|A,F) \\
&\approx \operatorname*{argmax}_{W_i} \sum_{U_i} P(U_i|A,F) P(W_i|U_i) P(A_i|W_i)
\end{aligned}
$$

where

$P(U_i|A,F)$ is posterior UT estimate using all available information

$P(W_i|U_i)$ is UT-specific language model

$P(A_i|W_i)$ is standard acoustic model

(assuming AM independent of UT)

Recognition with Discourse Grammar: 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.

Cheating Error Rates on Specific Utterance Types

Utterance Type	WER	Oracle WER	Improvement with Oracle
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%

What kind of errors did Cheating UT fix?

Original recognition:

```
REF:   DO YOU HAVE    SMOG LIKE   THEY DO     in ...
HYP:   ** *** YOU'RE NOT   SMOKE I     COULD in ...
```

Know that it's a question:

```
REF:   do you have SMOG LIKE THEY  DO       in ...
HYP:   do you have **** **** SMOKE ACTIVE in ...
```

Summary: Utterance Type Detection Accuracy

Chance	**35%**
Prosodic (HMM) + Discourse Bigrams	**45%**
Prosodic (Decision Tree) + Discourse Bigrams	**50%** *
Words (2500-best-list) + Discourse Bigrams	**64.6%**
Prosodic (DT) + Words (2500) + Discourse Bigrams	**65.0%**
Words (Cheating correct) + Discourse Bigrams	**71%**
Human Labelers	**83%**

* Single, preliminary decision tree

What We Could Have Done Better

- Better Discourse Grammar (modeling turn-changes, overlapping turns)
- Better way to combine decision tree models
- Better clustering of Utterance Types
- Better alignments on training data
- Use UT-detector to relabel corpus and retrain
- ME model to combine Discourse Grammar and Prosodic features

What We Learned

- Start meeting early: never too much preparation
- Tell everybody what you are doing
- Pick good people

Conclusions

- Utterance Type Detection
- First step toward Automatic Discourse Understanding
- Prosodic knowledge in recognition
- No recognition win; better recognition on tasks with many questions?

Part IV: Future Work

Extensions: Above and Below the Statement

Conversational Analysis

Sequences of utterances:

- Question/answer pairs
- Genres such as narrative

Internal structure of utterance type:

- Given/new structure of statements

Utterance Types in Context

Two areas analyzed:

1. Answers not explicitly marked in data
2. Statements in context within turns and across turns

Message: Not all statements are alike.

Context (both within and across turns) correlates with sentence length.

Context Affects Statement Length

Before		After	Length
Other statement	(s)	Other statement	4.33
Other question	(s)	Anything	8.06
Other backchannel	(s)	Self statement	10.42
Other statement	(s)	Other backchannel	10.50
Other backchannel	(s)	Self unfinished	10.57
Other backchannel	(s)	Other backchannel	12.94
Self unfinished	(s)	Other backchannel	13.07

Statement Grouping

Answers: Following question or marked

Short: Following other's statement; following own backchannel

Medium: Following own description statement

Long: Preceding other's backchannel; preceding or following own viewpoint statement

Looking Inside Utterances

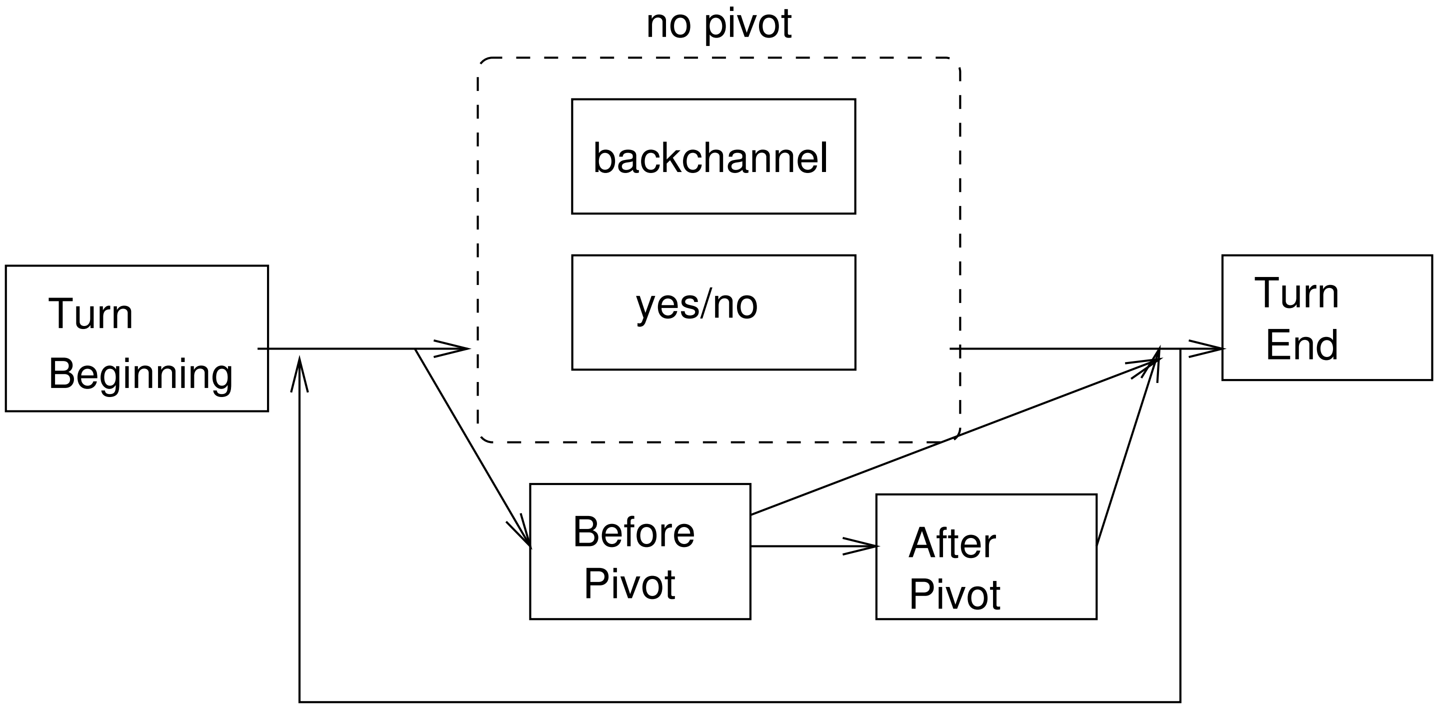
Given/new: Sentences are not uniform from beginning to end

- Given information tends to occur early
- New information tends to occur late

We can define a pivot based on the verb

```
- i've [PVT] voted in every major election
             since i turned twenty-one
- oh that's [PVT] great
```

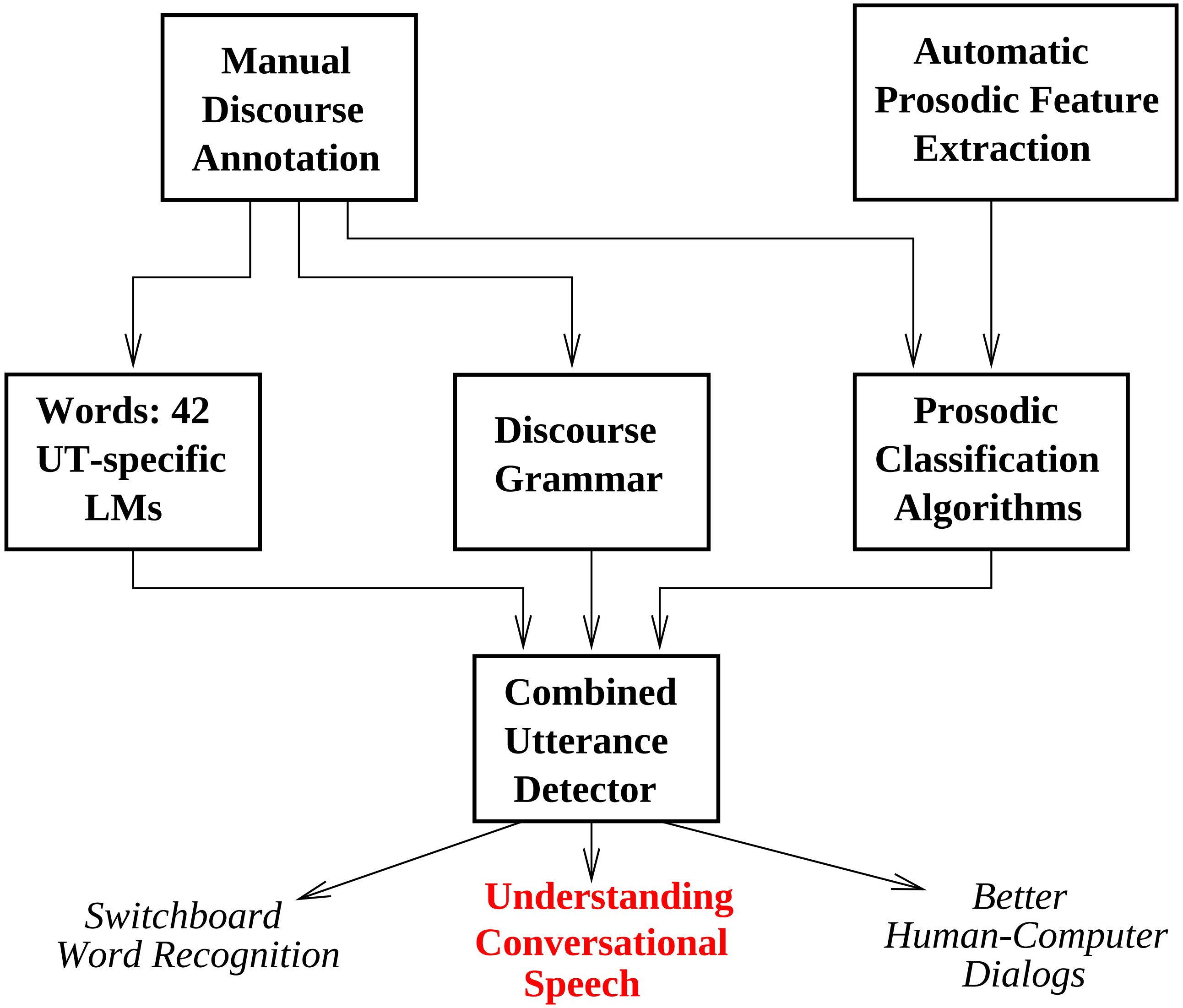
Given-New Structure of Utterances



Pivot Models Improve WER on Some UT's

			Improv in WACC from Baseline			
Train	WER	PPL	Stmt	Opinion	BackCh	Aband
Baseline	41.2	74.81				
Cheating	40.8	70.78	+0.71	+0.44	+7.72	+8.34
Pivot Algorithm	42.0	85	-0.22	-2.60	-1.11	-10.08
Pivot Automatic	41.7	80	+0.17	-1.16	-0.83	-7.89
Pivot Auto Interp	41.0	76	+1.00	-0.39	0.00	-5.25

Where are we?



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

Discourse Research
Pre-During-Post-WS'97
Carol Van Ess-Dykema (DoD)
August 21, 1997

Clarity Project

Carnegie Mellon University
MITRE Corporation
Department of Defense

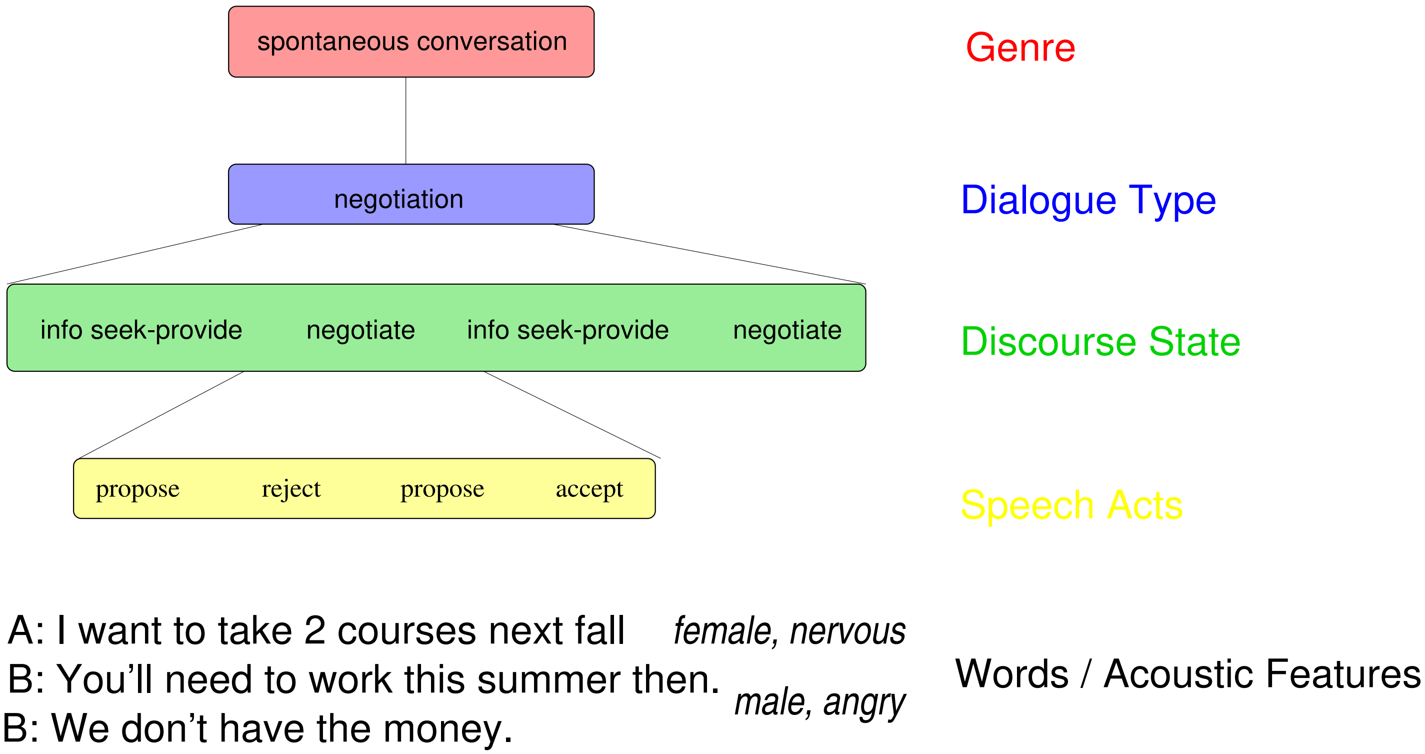
Clarity

Database is Spanish CallHome.

Different levels of knowledge to improve automated speech/transcript understanding

- LVCSR
- detection of prosodic features
- automated labeling of different layers of discourse

Understanding Spontaneous Dialogue



The Summer Workshop and Clarity (1)

- gaining experience on SWBD
- using the database to bootstrap initial development
- reusing software and labeling schemes
- work of various members of Clarity within/close to the workshop

The Summer Workshop and Clarity (2)

Interesting studies across corpora might be possible

- compare cross-lingual – CallHome Spanish / English
- compare cross-style – SWBD / CallHome English

Conclusions

Clarity will be a project that

- develops state-of-the-art LVCSR on CallHome Spanish
- fosters research in empirical methods for dialog processing
- focuses on language understanding for spontaneous and unrestricted speech
- integrates different levels of automated speech understanding in one system

Other Potential Uses for Discourse-Tagged SWBD

LDC Release of Data

Speech Recognition: • local caching

- immediate word recycling
- cross-speaker LMs
- conversational-level caching

Natural Language Processing: • discourse-robust ellipsis and anaphora resolution

- research on conversation structure
- subtopic identification
- use of discourse markers (cue words)
- models of disfluencies