Summarization of Conversational Multi-Party Speech

Michel Galley
Advisor: Kathleen McKeown

Columbia University
Department of Computer Science

AAAI/SIGART Doctoral Consortium
July 16-17, 2006

Why meeting summarization?

- Transcriptions provide poor information access
  - long meeting transcriptions (avg. 15’000 words, 1h.)
  - raw formatting: no sections, paragraphs, etc.
  - useful information is scattered across meeting (not only at the beginning)

	record
speaker A	i mean so i think pairwise relationships are pretty easy
speaker B	mm-hmm
speaker A	you know source destination relations .. are there other sorts of things that might we might want to record
speaker C	it's useful to know that that relationship
speaker A	i think that fits in well with the whole meeting map mapping meetings concept is that's another way of looking at looking at it
speaker C	interesting
speaker A	so are there anything other than pairwise
speaker C	oh well yeah you could have people who are all part of the same football league or uh or chess club or -

Introduction

Why meeting summarization?

- Transcriptions are hard to read
  - speech errors, hesitations, etc.
  - content-poor conversational expressions: “you know”, “I mean”, “sort of”, “kind of”, etc.

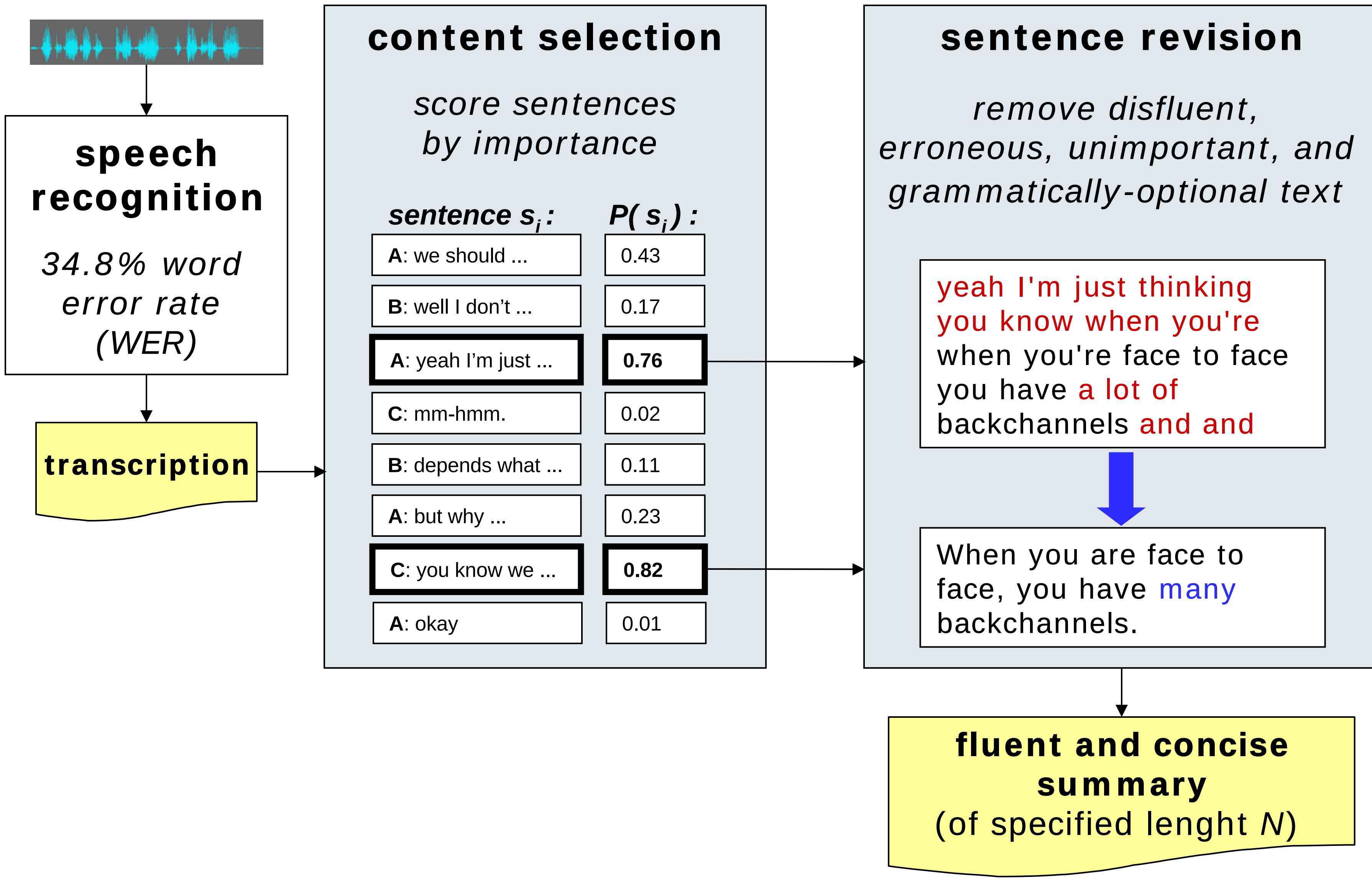
well **i ju- i was just thinking** with reference to **uh** things **that have -** that bear on the content or the status relations would be the things .. **without being exhaustive by any means** but just like i said if there's a **k-** a certain topic that comes up in the meeting and that knowing their relationship will clarify it or .. if there's a certain dynamic that comes up so **i mean** a person is asked a whole bunch of questions more than you'd usually think they'd be asked and **it turns out** it's because he's being prepared for a job interview **or something like that** then it's useful to know **that -** that relationship.

filler

self repair

content poor phrases

Two main problems: selection and revision

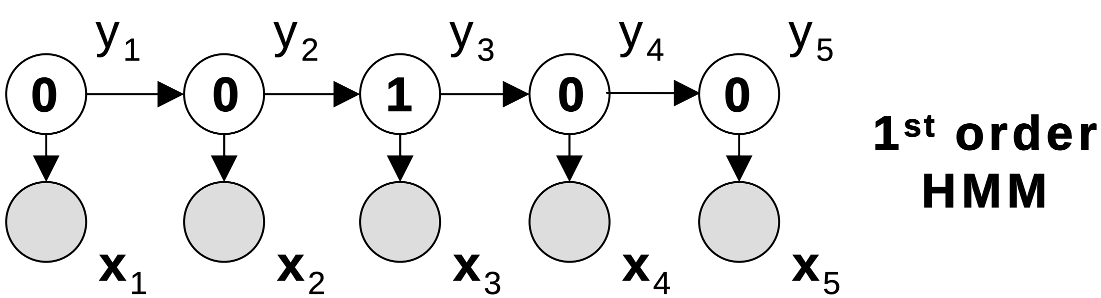


Outline

- **Content selection**
  - **Previous work**
  - **Research objectives**
  - **Approach overview**
  - **Open questions**
- Sentence revision
  - Previous work
  - Research objectives
  - Framework
  - Open questions

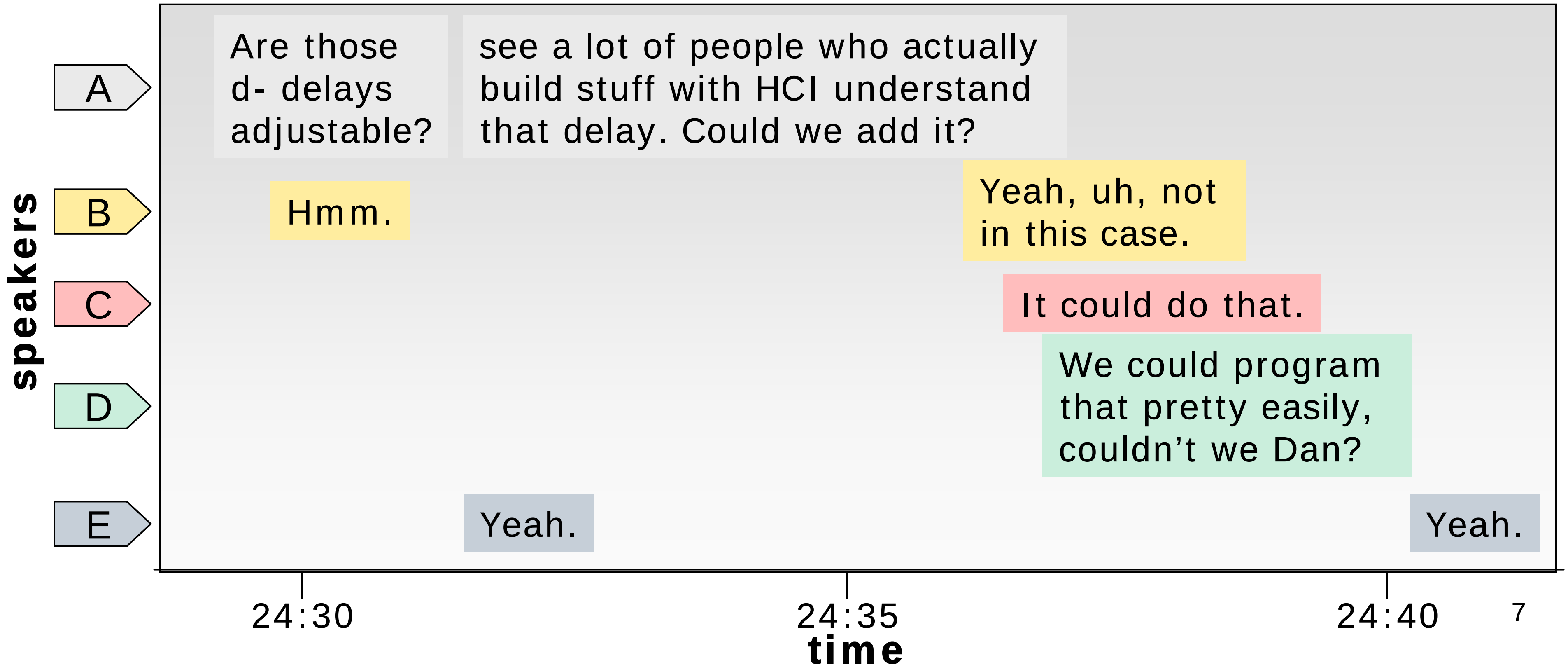
Content selection
Previous work

- Approaches
  - Extensive previous work: trainable, knowledge rich, IR-based, discourse-based (overview: [Mani and Maybury, 1999])
- Trainable summarizers
  - Binary classification at sentence level (Naive Bayes [Kupiec et al., 1995], etc.)
- Sequence classifiers
  - Markov models (e.g., HMM) [Conroy et al., 2005; Maskey and Hirschberg, 2006]
  - Best performing system in recent NIST summarization evaluations.
  - Well suited to written texts (sentences are linearly sequenced).



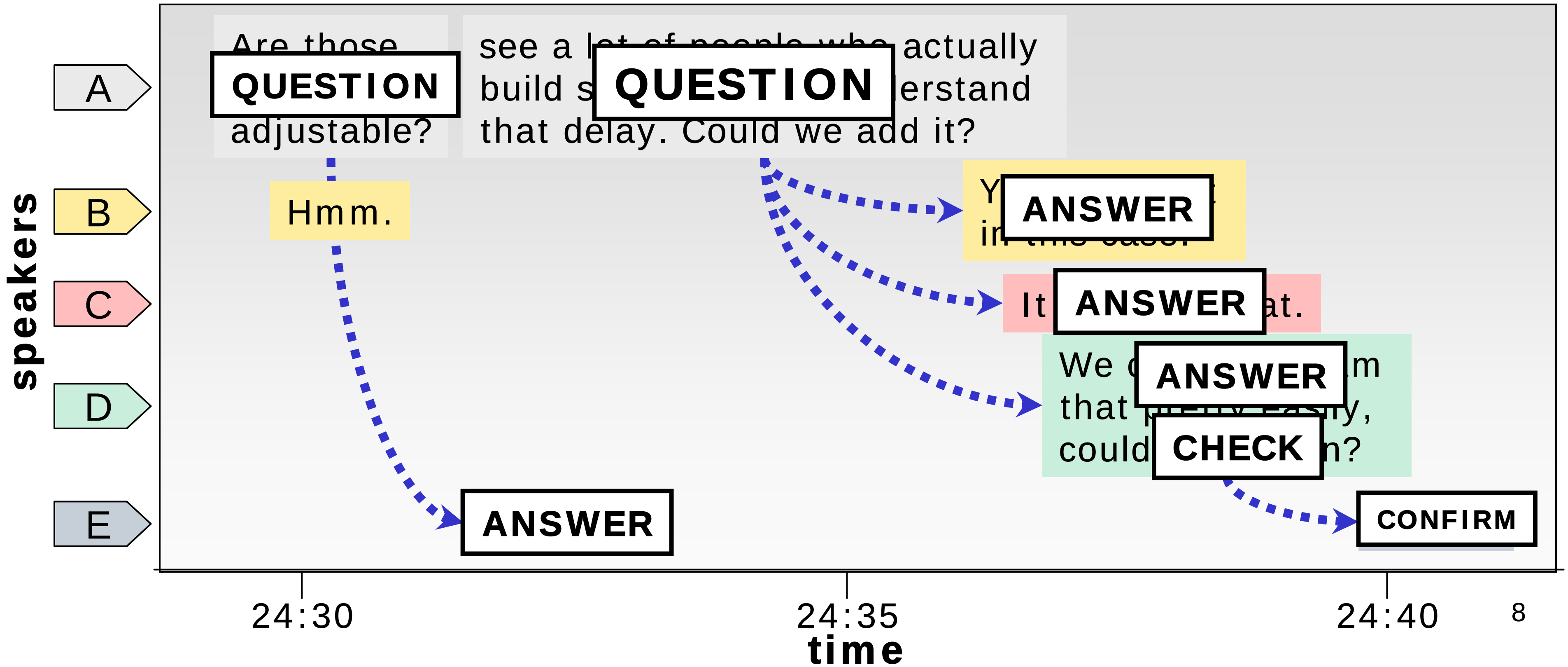
Research objectives

- Model selection for sequence classifiers
  - dependency structure, latent variables, network semantics (directed or undirected)
  - models that account for multi-party interaction (3+ speakers, overlapping speech)



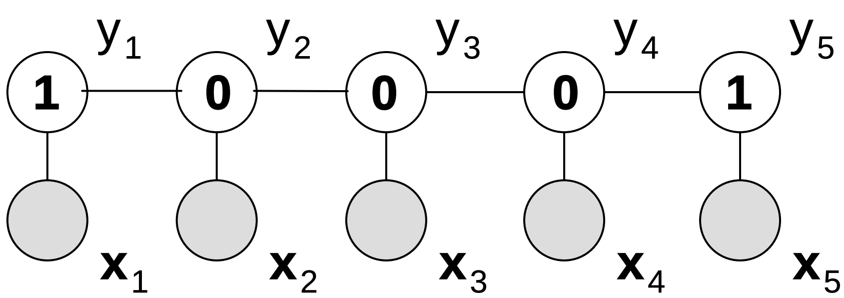
Research objectives

- Non-local structure
  - model interaction between arbitrary-distant sentences (e.g., QUESTION-ANSWER, OFFER-ACCEPT, CHECK-CONFIRM)



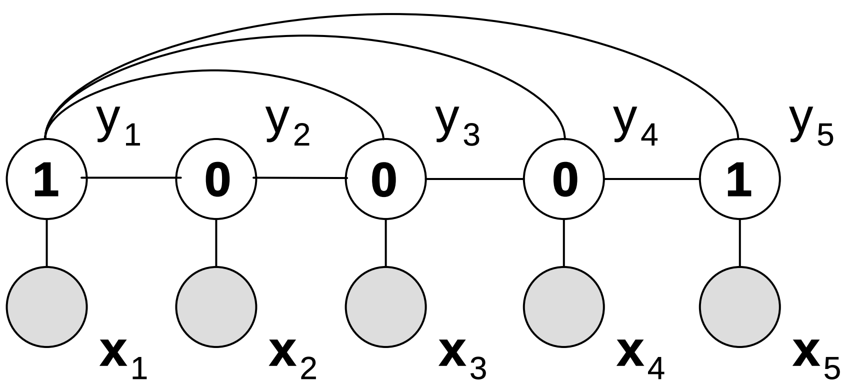
Content selection

Model structure: linear vs. skip-chain

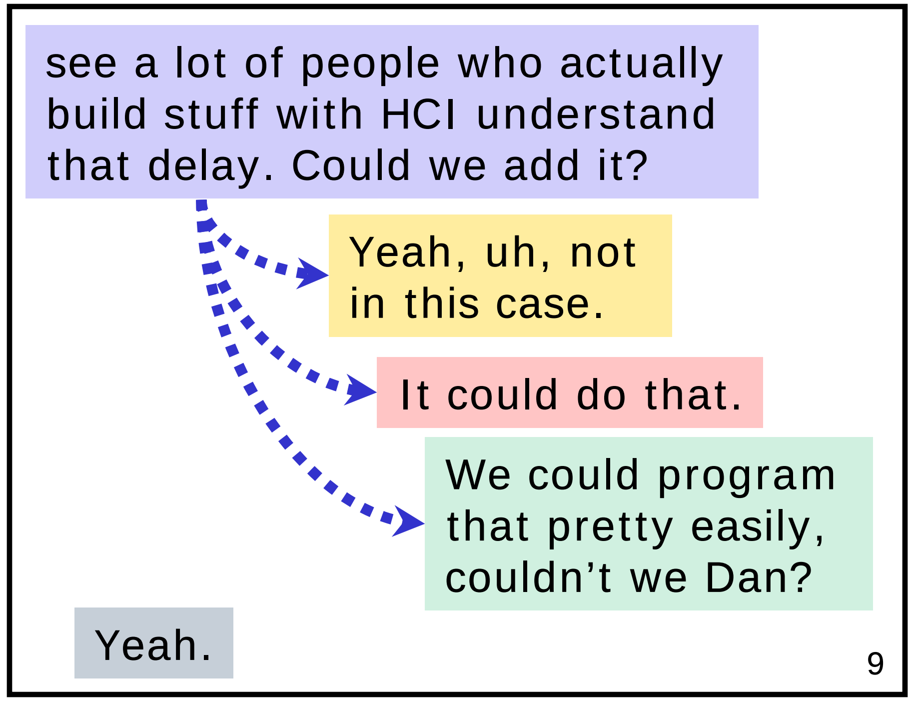


linear-chain

see a lot of people who actually build stuff with HCI understand that delay.
Yeah.
Yeah, uh, not in this case.
It could do that.
We could program that pretty easily, couldn't we Dan?

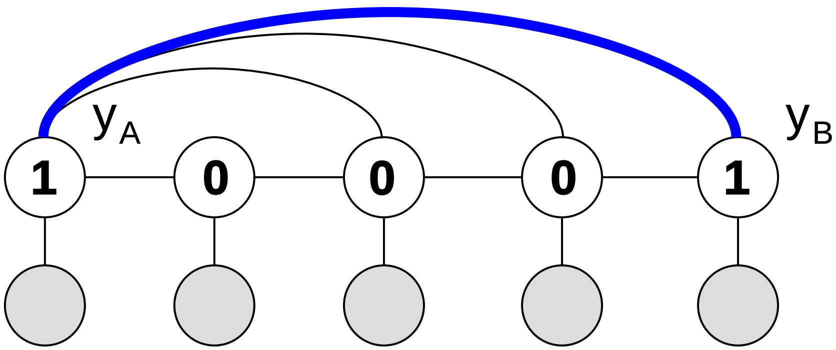


"skip-chain"



Content selection
Model assessment: linear vs. skip-chain

Are dynamic conditioning variables really useful?



skip-chain edges

	$y_B=1$	$y_B=0$
$y_A=1$	6’792	2’191
$y_A=0$	1’479	121’591

contingency tables

chi-sq test very significant ($p<.001$)

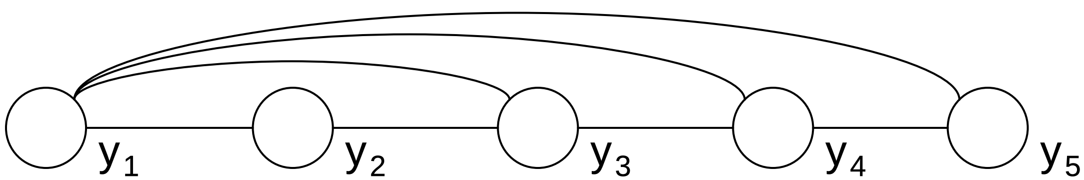
Content selection
Approach Overview

Model structure inference

[Galley, McKeown, Hirschberg, Shriberg; ACL-2004]

- Identify speaker-addressee (SA) links, as between QUESTION-ANSWER, OFFER-ACCEPT:

  given sentence B (e.g., ANSWER), find corresponding sentence A (e.g., QUESTION).
- Rank candidate A parts with log-linear model (0.92 accuracy).



Content selection with inferred graphical model

[Galley; EMNLP-2006]

- Classification with sequential and non-sequential classifiers.
- Inference with skip-chain conditional Markov random fields (CRFs) and Bayes nets (BNs).
- CRFs achieved best results.

Content selection
Ranking sentences

Three ranking functions to extract an n% summary:

- Binary predictions
  - Only include positive predictions, i.e. $P(y_i = 1|\ldots) \geq .5$ (trim summary if too long)
- Class posteriors for BNs
  - Ignore predictions; rank utterances by $P(y_i = 1| \ldots)$
- Class posteriors for CRFs
  - Problem with CRFs: sum of potentials have no probabilistic interpretation, i.e. can't be used to estimate $P(y_i | \ldots)$.
  - A solution: since CRF and BN are parameterized with the same feature functions, we can:
    1. train and decode optimal sequence $(\hat{y}_1,\ldots,\hat{y}_T)$ with CRF
    2. estimate $P(y_i = 1| \hat{y}_1, \ldots, \hat{y}_{i-1}, \ldots)$ with BN model

Content selection
Results

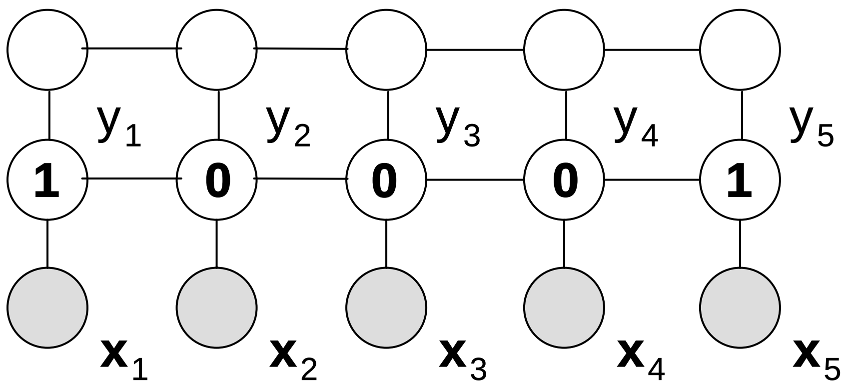
- CRFs outperform equivalent directed models (Bayes nets)
- Skip-chain CRFs outperform linear-chain models
- Ranking by posteriors outperforms 0/1 predictions

		Markov order			
Model	Ranking	k=0	*k*=1	*k*=2	*k*=3
linear-chain BN	0/1 predictions		.241	.267	.267
linear-chain BN	posteriors	.511	.512	.519	.525
skip-chain BN	posteriors		.543	.549	.542
linear-chain CRF	0/1 predictions		.326	.36	.348
linear-chain CRF	posteriors	.511	.53	.548	.54
skip-chain CRF	posteriors		.541	.554	.559

average F-score

Content selection
Open questions

- Do extra hidden variables interact with observation or state variables?
  - topic variables [Barzilay and Lee, 2004]
  - dialog acts (DA) variables, e.g. $\in$ {STATEMENT, Y/N-QUESTION, CHECK, ...}



- Perform joint inference?
  - speaker-addressee identification and content selection as a joint learning problem (instead of two-step approach)

Outline

- Utterance selection:
  - Previous work
  - Research objectives
  - Approach overview
  - Open questions
- **Utterance revision:**
  - **Previous work**
  - **Research objectives**
  - **Approach**
  - **Open questions**

Previous work: two main categories

- Word-based models [Banko et al., 2000]
  - Word deletion models: $P_{\text{delete}}(\text{“not”}) < P_{\text{delete}}(\text{“also”})$
  - Works well with short sentences (e.g., headlines)
  - No direct way of preserving grammaticality: produces ill-formed sentences on long inputs

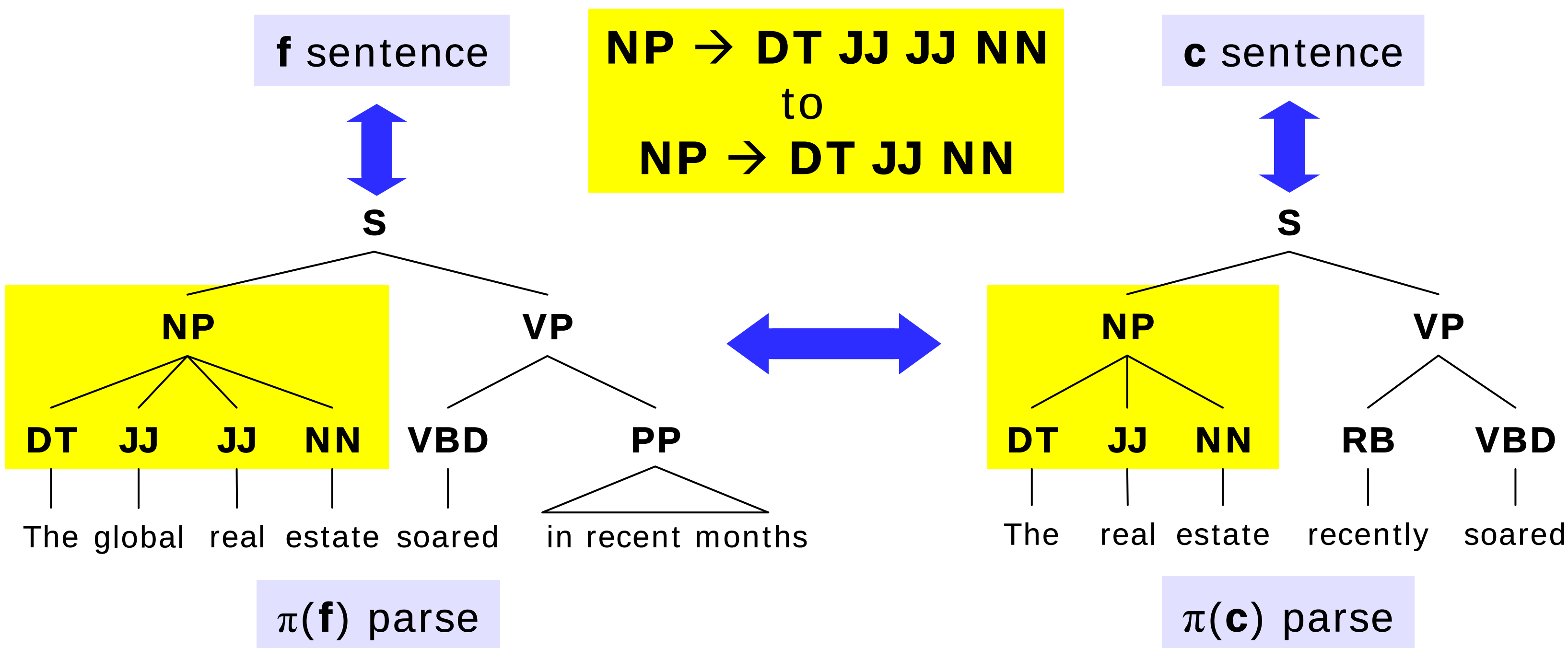
- Syntax-based models [Knight and Marcu, 2000; Turner and Charniak, 2005]
  - Transform syntactic analysis of **f** into a reduced one
  - Output presumably more grammatical
  - Word deletion probabilities not lexicalized: $P_{\text{delete}}(\text{“not”}) = P_{\text{delete}}(\text{“also”})$ (since both adverbs)

Research objectives

- Fully trainable sentence revision model
  - transformational model mapping a full sentence $\mathbf{f}=(f_1,\ldots,f_n)$ to a subsequence $\mathbf{c}=(c_1,\ldots,c_m)$: $P(\mathbf{c}|\mathbf{f})$
  - fully trainable from ($\mathbf{c}$,$\mathbf{f}$) pairs
- Syntax-driven revision rules
  - syntactic transformation rules to map from **c** to **f**, e.g. [Det Adj Noun] → [Det Noun]
- Effective estimation of rule probabilities
  - factorization of rule probabilities: computationally and linguistically motivated
  - Lexicalized compression models, e.g. more likely to delete “also” than “not”
  - Integration of any arbitrary feature: IR (TF.IDF), acoustic, etc.

Framework

- Synchronous grammars
  - model the **f** ←→ **c** transformation indirectly through their respective syntactic analysis
  - many resources (e.g. parsers) to get **f**→π(**f**) and **c**→π(**c**)
  - easier to define grammaticality and meaning preserving operations on **context free grammar (CFG)** productions

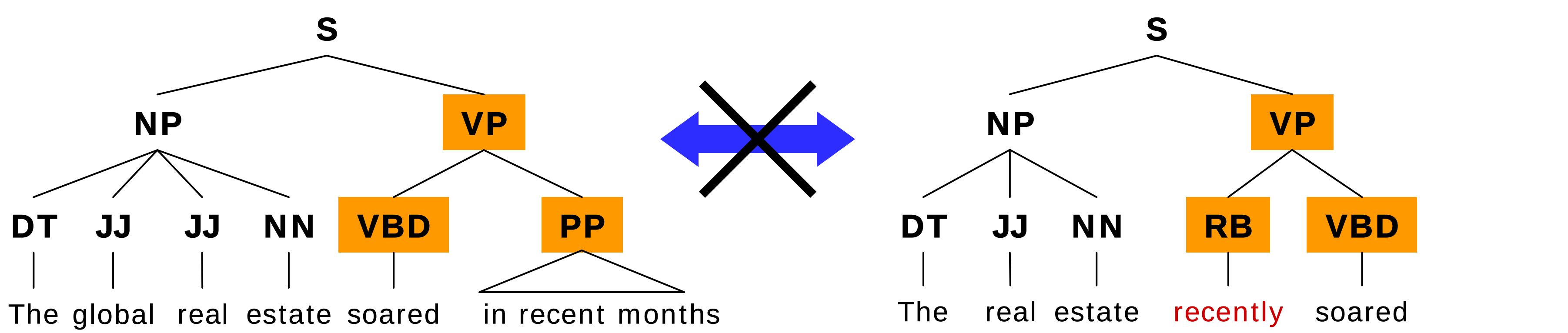


Sentence revision

Rule extraction

- Extracting grammar rules from sentence pairs
  - Previous approaches:
    - assume that $\pi(\mathbf{c})$ is a trimmed version of $\pi(\mathbf{f})$, i.e., that **c** is a subsequence of **f**
    - assumption almost always incorrect → low coverage (e.g., can only use 2.7% of the Ziff-Davis parallel corpus for training [Knight and Marcu, 2000])

Tree pair is discarded because of one word insertion ("recently"), though we could try to learn to compress "the global real estate":

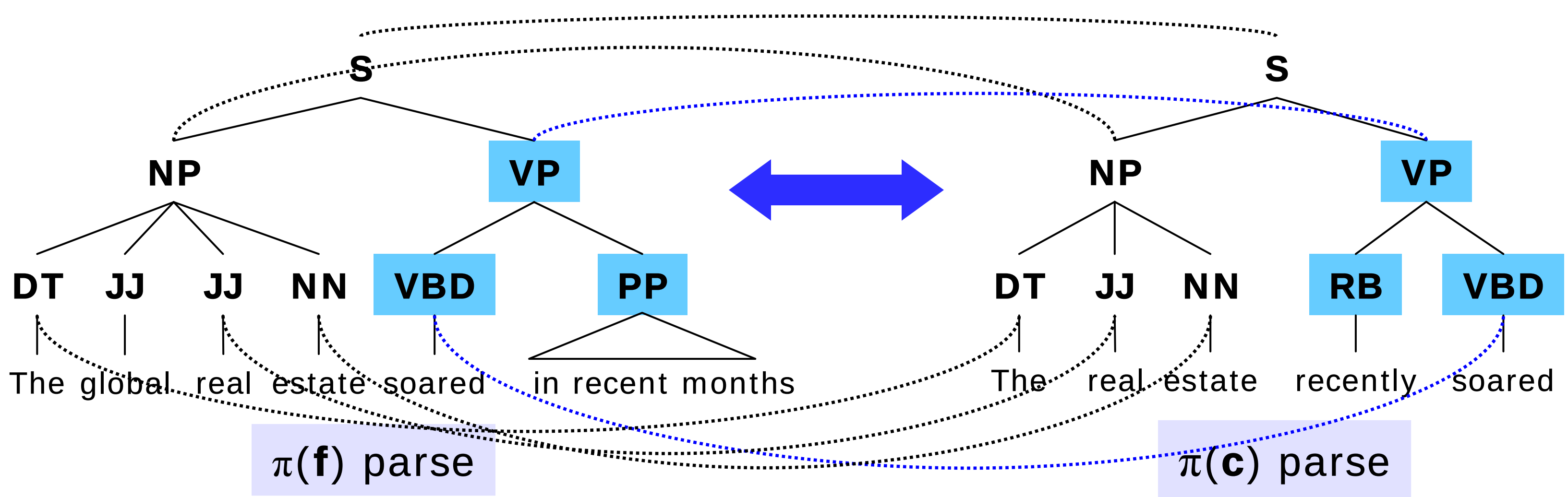


$\pi(\mathbf{f})$ parse

$\pi(\mathbf{c})$ parse

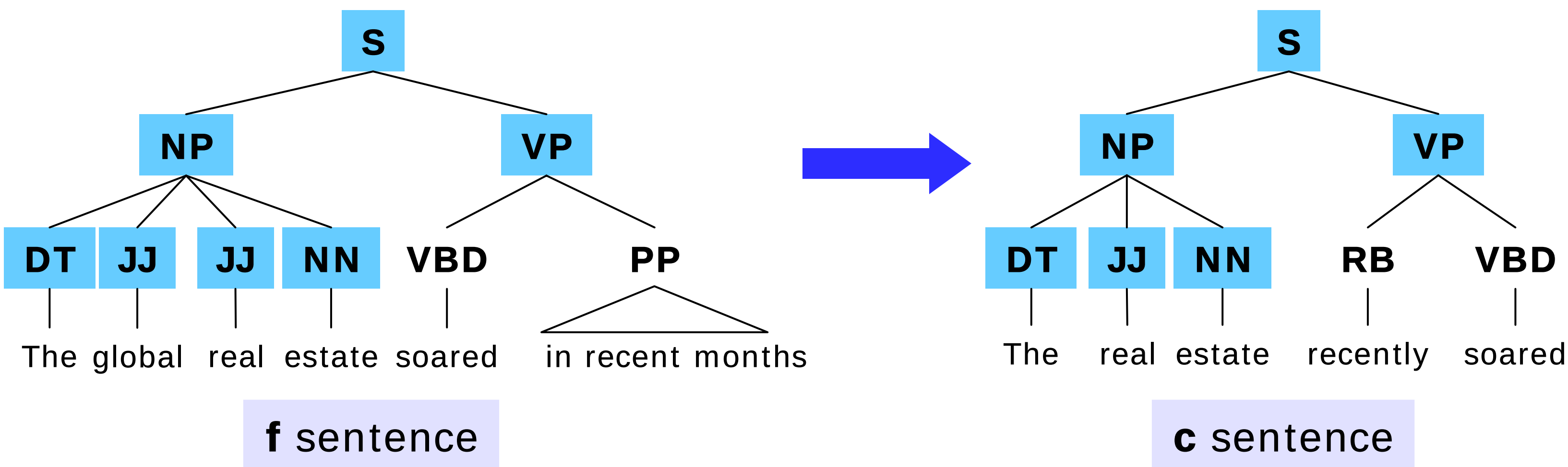
Sentence revision
Rule extraction

- Extracting grammar rules from sentence pairs
  - Proposed approach:
    - tree-to-tree alignments (e.g., min. tree edit distance)
    - bijection between tree alignment and grammar rules (synchronous tree substitution grammar)



Full generative story: advantages

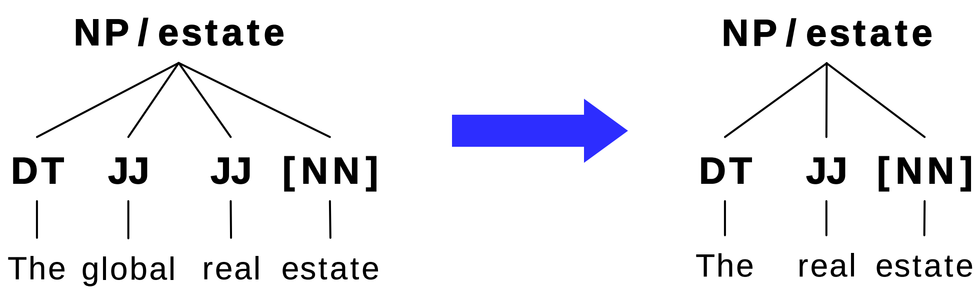
- Increased data usage:
  - can align many tree pairs (»2.7%) → more counts for CFG compression rules
  - in practice, most rules are CFG compressions:



- Richer revision rules (non CFG):
  - tree-to-tree rewrite rules: covers many deletion not possible with , such as deleting “the spokesman said” in “**S**, the spokesman said”.

Effective parameterization of *P*(c|f)

- Generative model *P*(c,f)
  - Make major independence assumptions (similar to [Collins, 1999])
  - Introduce bi-lexical dependencies:
    - *"real"* modifying *"estate"* : low JJ-deletion probability
    - *"global"* modifying *"estate"* : higher JJ-deletion probability



- Discriminative model *P*(c|f)
  - any arbitrary feature (TF.IDF, LM score, etc.) weighted with, e.g., SVM, perceptron, linear regression
  - global features computed in post processing stage (n-best re-ranking)

Sentence revision

Open questions

- Synchronous grammars
  - how to best factorize rule probabilities? prevent data sparseness while avoiding unreasonable independence assumptions
- Text-to-text generation
  - not just deletions, but insertions and substitutions (e.g., “a lot of” → “many”, etc.)
- Integration with content selection
  - how to balance compression level in content selection and sentence revision
  - choices made in sentence revision can affect selection

Progress to date / Plan for completion

- Content selection
  - skip-chain CRFs for content selection **[completed]**
  - joint inference: skip-chain identification and content selection **[future]**
- Sentence revision
  - Corpus of 5’000 (sentence, revision) pairs **[current]**
  - Syntactic compression models **[current]**
  - Decoding most likely compressed sentence **[current]**
  - Discriminative re-ranking with arbitrary features (TF.IDF, etc.) **[future]**
  - From compression to revision **[future]**

Contributions

- Content selection
  - model of long-distance relationships (speaker-addressee)
  - use of those relationships for better content selection
- Sentence revision
  - alignment between any string pair (**f**,**c**):
    - better corpus coverage (more data)
    - more complex revision operations
  - empirical evaluation of different rule parameterization (lexicalized or not, etc.)
  - re-ranking framework where any arbitrary summarization feature can be added (e.g., TF.IDF)

Acknowledgements

- Thesis committee
  - Julia Hirschberg (chair), Columbia U.
  - Kathleen McKeown (advisor), Columbia U.
  - Owen Rambow, Columbia U.
  - Daniel Ellis, Columbia U.
  - Michael Collins, MIT
- Co-authors
  - Elizabeth Shriberg
  - Eric Fosler-Lussier
  - Hongyan Jing
  - Kevin Knight
  - Daniel Marcu
  - Mark Hopkins

QUESTIONS