Attention Shifting for Parsing Speech

Keith Hall and Mark Johnson

Brown University

ACL 2004

July 22, 2004

Attention Shifting

- Iterative best-first word-lattice parsing algorithm
- Posits a complete syntactic analyses for each path of a word-lattice
- Goals of Attention Shifting
  - Improve accuracy of best-first parsing on word-lattices
    (Oracle Word Error Rate)
  - Improve efficiency of word-lattice parsing
    (Number of parser operations)
  - Improve syntactic language modeling based on multi-stage parsing
    (Word Error Rate)
- Inspired by edge demeriting for efficient parsing
  Blaheta & Charniak demeriting (ACL99)

Outline

- Syntactic language modeling
- Word-lattice parsing
- Multi-stage best-first parsing

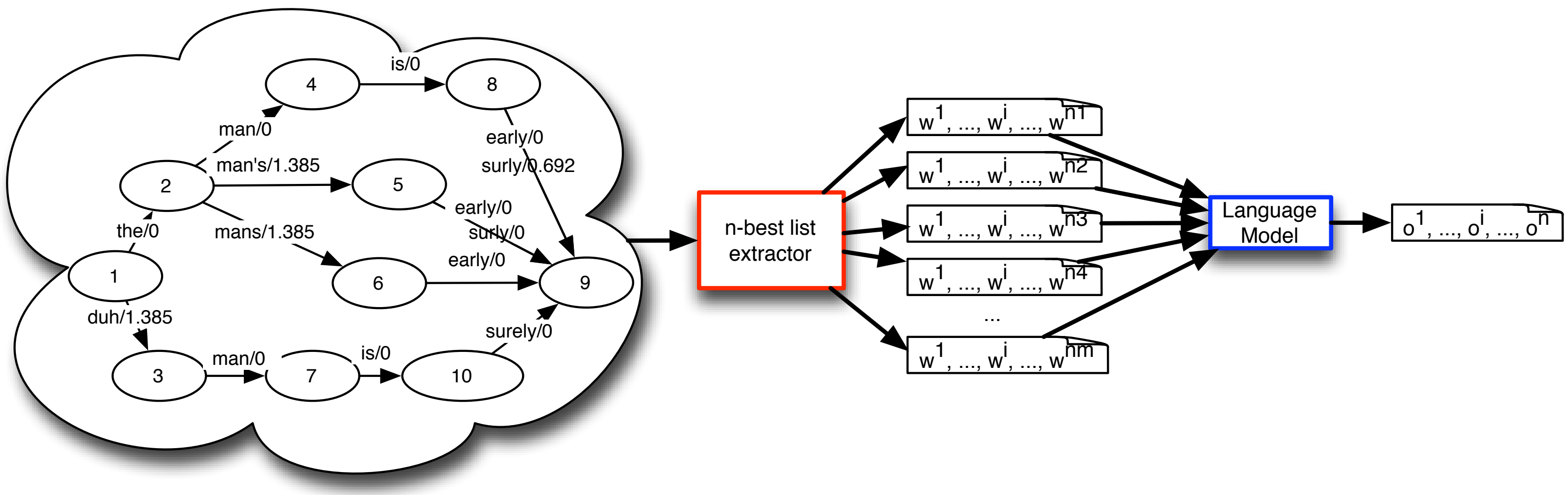
Noisy Channel

$$P(A,W) = \underbrace{P(A|W)}_{\text{Noise Model}}\underbrace{P(W)}_{\text{Language Model}}$$

- Speech recognition: Noise model = Acoustic model

$$\arg\max_{W} P(W|A) = \arg\max_{W} P(A,W)$$

Syntactic Language Modeling

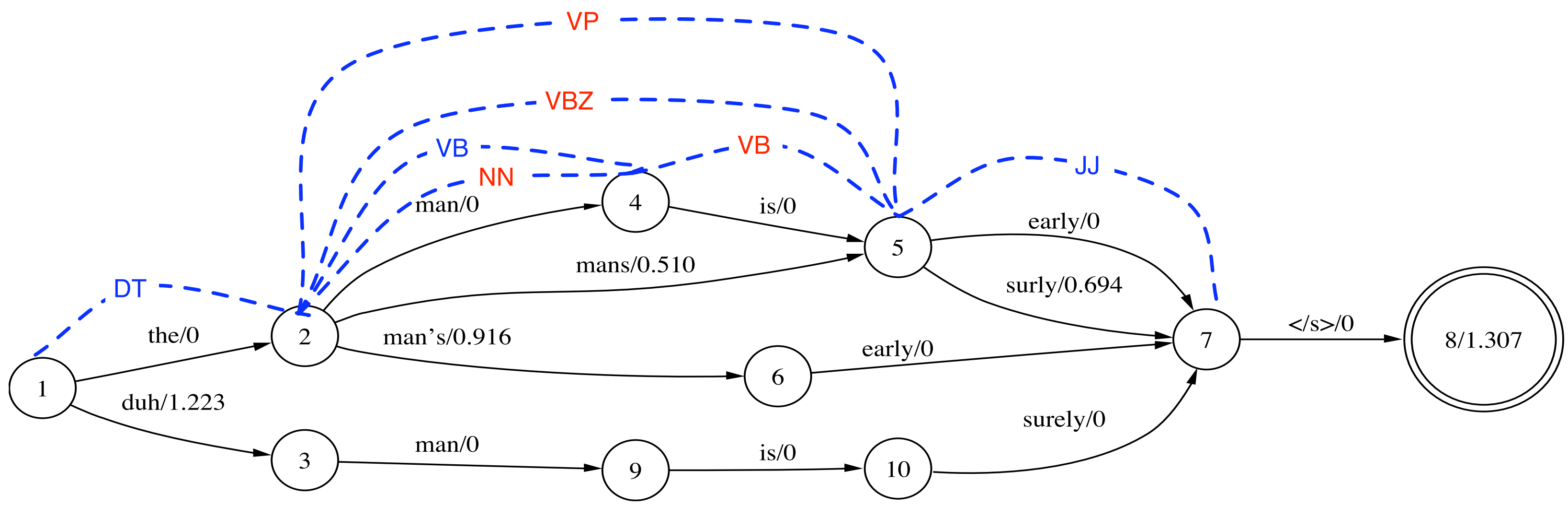


- Adding syntactic information to context (conditioning information)

  $P(W) = \prod_1^k P(w_i|\pi(w_k, \ldots, w_1))$

- n-best reranking
  - Select n-best strings using some model (trigram)
  - Process each string independently
  - Select string with highest $P(A, W)$
- Charniak (ACL01), Chelba & Jelinek (CS&L00,ACL02), Roark (CL01)

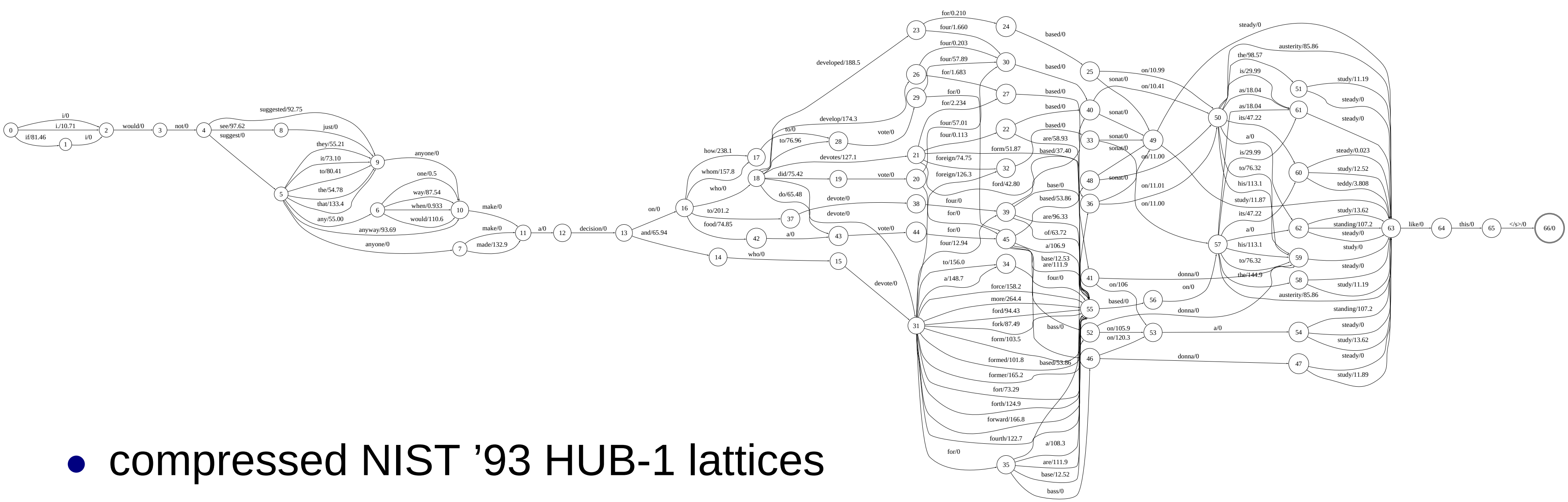
Parsing word-lattice



- Compress lattice with Weighted FSM determinization and minimization (Mohri, Pereira, & Riley CS&L02)
- Use compressed word-lattice graph as the parse *chart*
- Structure sharing due to compressed lattice
  VP → NN VB covers string *man is*
  VP → VBZ covers string *mans*

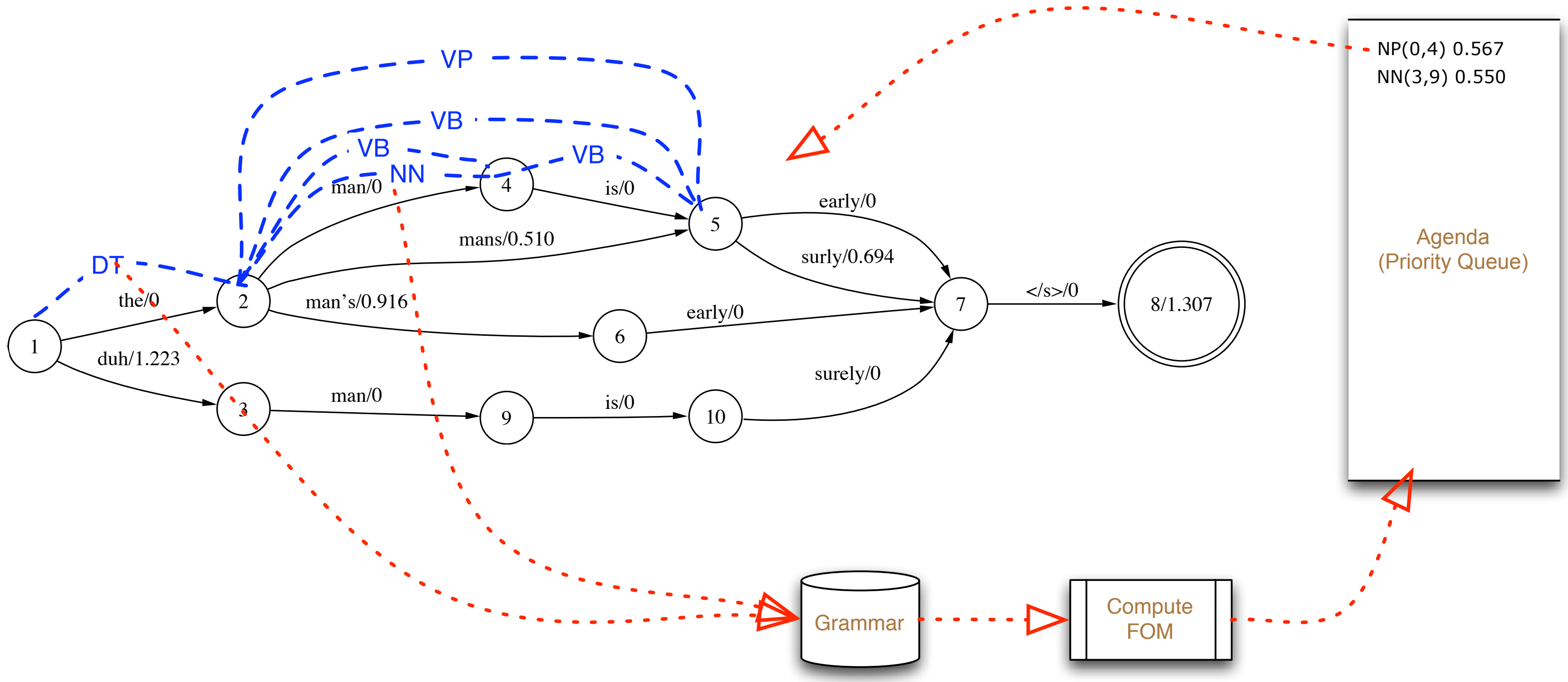
Word-lattice example

- I WOULD NOT SUGGEST ANYONE MAKE A DECISION ON WHO TO VOTE FOR BASED ON A STUDY LIKE THIS
  (160 arcs, 72 nodes)



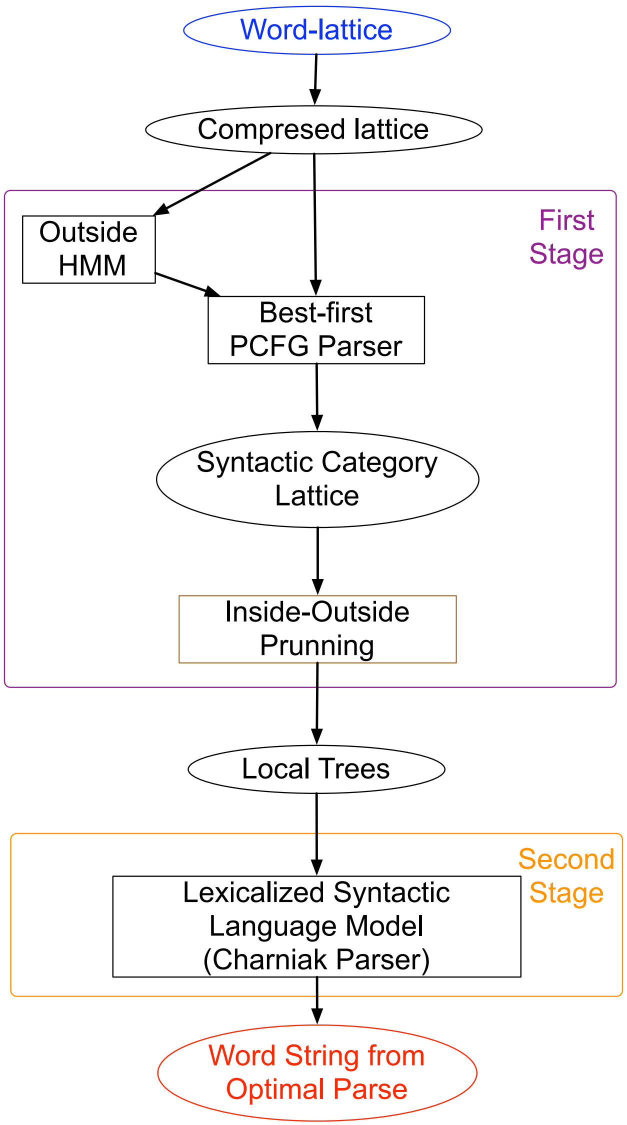
- compressed NIST ’93 HUB-1 lattices
  – average of 800 arcs/lattice (max 15000 arcs)
  – average of 100 nodes/lattice (max 500 nodes)

Best-first Word-lattice Parsing



- Bottom-up best-first PCFG parser
- Stack-based search technique based on figure-of-merit
- Attempts to work on “likely” parts of the chart
- Ideal figure-of-merit: $P(edge) =$ inside$(edge) *$ outside$(edge)$
  details in (Hall & Johnson ASRU03)

Word-lattice Parsing

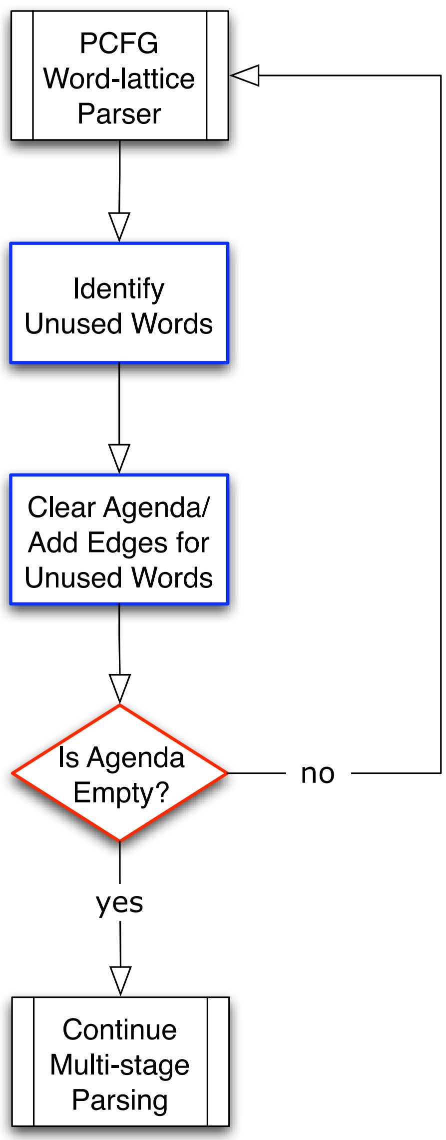


- First stage: best-first bottom-up PCFG parser
- Second stage: Charniak Parser Language Model (Charniak ACL01)
- Parsing from lattice allows structure sharing
- Combines search for candidate lattice paths with search for candidate parses

Multi-stage Deficiency

- First-stage PCFG parser selects parses for a subset of word-lattice paths
- Lexicalized syntactic analysis not performed on all of the word-lattice
- Covering entire word-lattice requires excessive over-parsing
  - 100X over-parsing produces forests too large for lexical-parser
  - additional pruning required, resulting in loss of lattice-paths
- Attention shifting algorithm addresses the coverage problem

Attention Shifting



- Iterative reparsing
  1. Perform best-first PCFG parsing (over-parse as with normal best-first parsing)
  2. Identify words not covered by a complete parse (unused word has 0 outside probability)
  3. Reset parse Agenda to contain unused words
  4. If Agenda $\neq \emptyset$ repeat
- Prune chart using inside/outside pruning
- At most $|A|$ iterations ($|A|$ = number of arcs)
- Forces *coverage* of word-lattice

Experimental Setup

- PCFG Parser trained on Penn WSJ Treebank f2-21,24
  (speech-normalization via Roark's normalization)
  - Generated at most 30k local-trees for second-stage parser
- Lexicalized parser: Charniak's Language Model Parser
  (Charniak ACL01)
  - trained on parsed BLLIP99 corpus (30 million words of WSJ)
  - BLLIP99 parsed using Charniak string parser trained on Penn WSJ

Evaluation

- Evaluation set: NIST '93 HUB-1
  - 213 utterances
  - Professional readers reading WSJ text
- Word-lattices evaluated on:
  - n-best word-lattices using Chelba A* decoder (50-best paths)
  - compressed acoustic word-lattices
- Metrics
  - Word-lattice accuracy (first-stage parser): Oracle Word Error Rate
  - Word-string accuracy (multi-stage parser): Word Error Rate
  - Efficiency: number of parser agenda operations

Results: n-best word-lattices

- Charniak parser run on each of the n-best strings (reranking)
  (4X over-parsing)
- n-best word-lattice: pruned acoustic word-lattices containing only n-best word-strings
- Oracle WER of n-best lattices: 7.75

Model	# edge pops	Oracle WER	WER
n–best (Charniak)	2.5 million	7.75	11.8
100x LatParse	3.4 million	8.18	12.0
10x AttShift	**564,895**	**7.78**	11.9

Results: Acoustic word-lattices

- Compressed acoustic lattices

Model	# edge pops	Oracle WER	WER
acoustic lats	N/A	3.26	N/A
100x LatParse	3.4 million	5.45	13.1
10x AttShift	**1.6 million**	**4.17**	13.1

Conclusion

- Attention shifting
  - Improves parsing efficiency
  - Increases first-stage accuracy (correcting for best-first search errors)
  - Does not improve multi-stage accuracy
- Pruning for second-stage parser constrains number of edges
- Useful for best-first word-lattices parsing