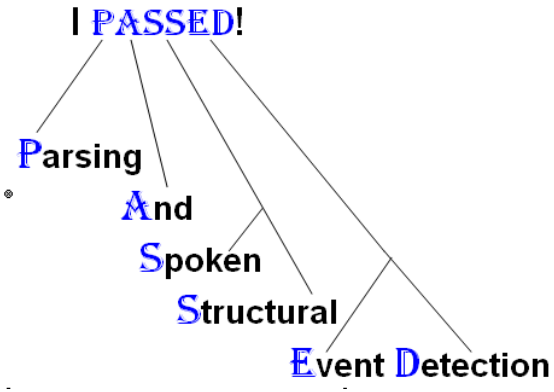
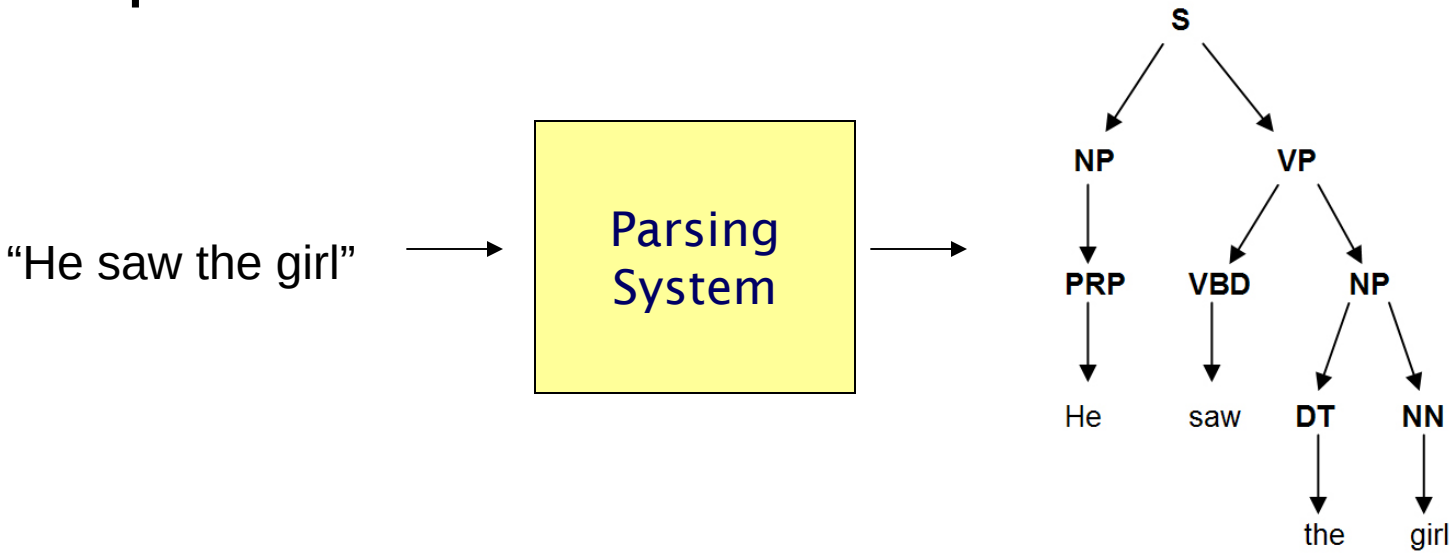
Hopkins CLSP 2005: Parsing and Structural Metadata in Speech

Where Parsing Meets Speech and Metadata Makes it Possible



The Parsing Problem

- Input: sentence $w_{1,m}$
- Output: parse tree
- Example:

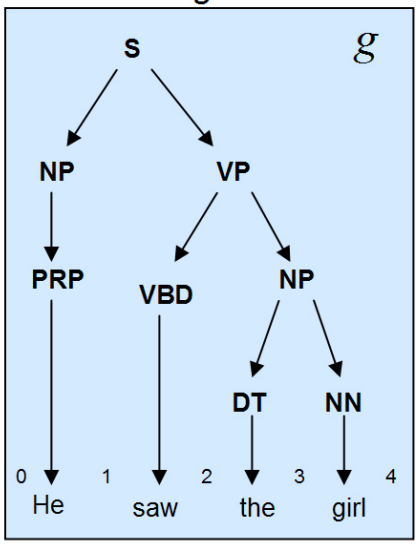


Parsing Metrics

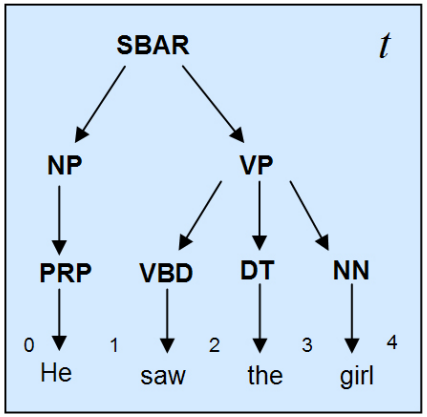
$$brackets(g) = \{S(0,4),\ \underline{NP(0,1)},\ \underline{VP(1,4)},\ NP(2,4)\}$$

$$brackets(t) = \{SBAR(0,4),\ \underline{NP(0,1)},\ \underline{VP(1,4)}\}$$

gold standard



evaluation tree



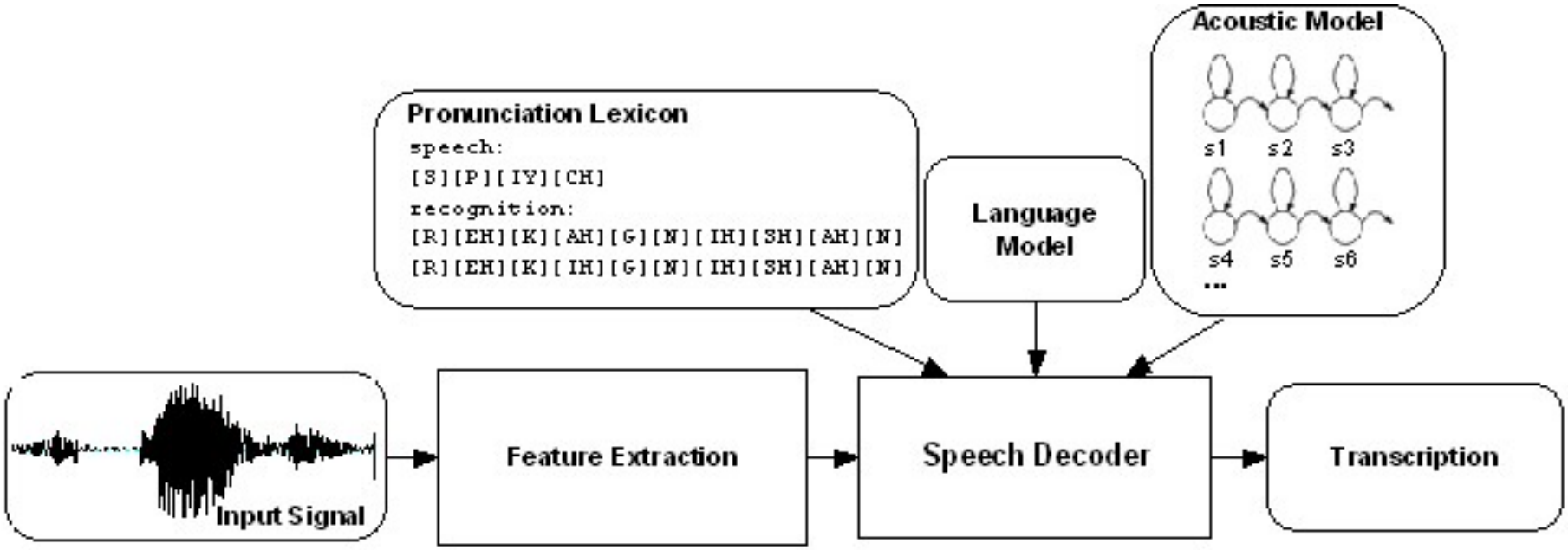
$$LP(t,g) = \frac{2}{3} = 66.66\%$$

$$LR(t,g) = \frac{2}{4} = 50.00\%$$

$$F_{meas}(t,\ g) = \frac{2 \cdot 66.66 \cdot 50}{66.66 + 50} = 57.14\%$$

State of the art on WSJ PTB is 91% F-measure with reranking parser.

Speech Recognition



ASR Output vs. Human Transcription

- **ASR Output**:

  {i} i kind of see that that you know perhaps you know we may need to like you're {get} close to the family environment and in {and} get down to the values of you know i mean no and {it's} money seems to be too big of an issue we would {wi- with with with} with with really was we would what's going on today

- **Human Transcription:**

  i i kind of see that that you know perhaps you know we may need to like get close to the family environment and and get down to the values of you know i mean uh it's money seems to be too big of an issue wi- with with with with with what's going on today

Human Transcription vs. Enriched Transcription

- **Human Transcription:**

  i i kind of see that that you know perhaps you know we may need to like get close to the family environment and and get down to the values of you know i mean uh it's money seems to be too big of an issue wi- with with with with with what's going on today

- **Enriched Human Transcription:**

  i i kind of see that that you know perhaps you know we may need to like get close to the family environment and and get down to the values of you know i mean uh it's money seems to be too big of an issue wi- with with with with with what's going on today

The Challenge of Parsing Speech

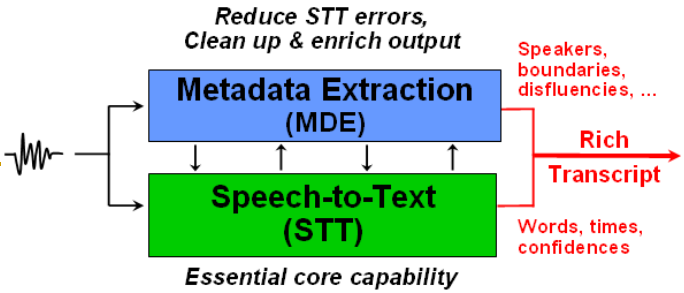
- There is a mismatch between ASR systems and statistical parsers:
  - Segments processed by an ASR system do not typically correspond to segments that statistical parsers normally work with.
  - ASR systems:
    - Produce long word strings without punctuation,
    - Word strings often contain errors (insertions, deletions, and substitutions),
    - Word strings contain phenomena that do not typically occur in textual sources (e.g., filled pauses, speech repairs).
  - Traditional parsers are text-based:
    - Don't use acoustic cues,
    - Process sentences not segments,
    - Process input without word errors,
    - Process textual input without spontaneous speech phenomena.

The Challenge of Spontaneous Speech

- Difficult for the recognizer
  - Acoustic challenges (fragments, filled pauses, coarticulation)
  - Language models do not currently model disfluencies adequately
- Recognition output is difficult for humans to read
- Recognition output is difficult for a parser
  - Parsing systems have a polynomial time complexity in the number of words; parsing conversation sides without segmentation is infeasible
  - Sentence boundaries are NOT provided and ASR segments are inappropriate
  - Utterances are different (planned on the fly) from written text
  - Much of spoken language is used for organizing the communication (e.g., “And so”).
  - Speech repairs are tough for many standard parsers.

Motivation for Rich Transcriptions

- Adding additional information to a transcription should:
  - **Aid downstream language processing (provide sentence boundaries, indicate structure of disfluencies for parsing)**
  - Improve readability to humans (adding punctuation, removing disfluencies) [e.g., MITLL readability experiments]
  - Improve ASR performance (e.g., feedback metadata information to recognizer to aid language models) [e.g., Work by Sebastien Coquoz, visiting ICSI from EPFL]



EARS Structural MDE Tasks

- **Sentence Unit (SU) detection:** find the sentence-like units **(*/*)** and their subtypes
- **Filler word detection:** filled pauses, discourse markers (e.g., **<you know>**), explicit editing terms
- **Edit word detection:** reparandum region of a speech repair (e.g., **[we were]** * I was fortunate)
- **Interruption point (IP) detection:** includes the IP inside edit disfluency and the point before filler words **(*)**

Each task has been evaluated on reference transcriptions (REF) and speech recognition output (STT).

An MDE Example

so we need but how do we get them out I say we have we set a string of charges that will root them out the back so t- the charges start at the front and just explode and blow a little something up but are really really loud and and marsupials have really good ears so that'll be real that'll really frighten them

Word Stream with Structural Metadata

- [so we need] * but how do we get them out /?
- I say [we have] * **we set** a string of charges that will root them out the back /.
- <so> [t-] * **the** charges start at the front and just explode and blow a little something up but are really really loud /.
- [and] * **and** marsupials have really good ears /.
- <so> [that'll be real] * **that'll really** frighten them /.

MDE Scoring Metric

- MDE scoring tool first aligns recognition words to reference transcripts and then maps metadata events
  - Error rate = # errors / **# reference events**
- An example of SU detection output:

Reference: w w | w w w w |

System: w w w | w w w |

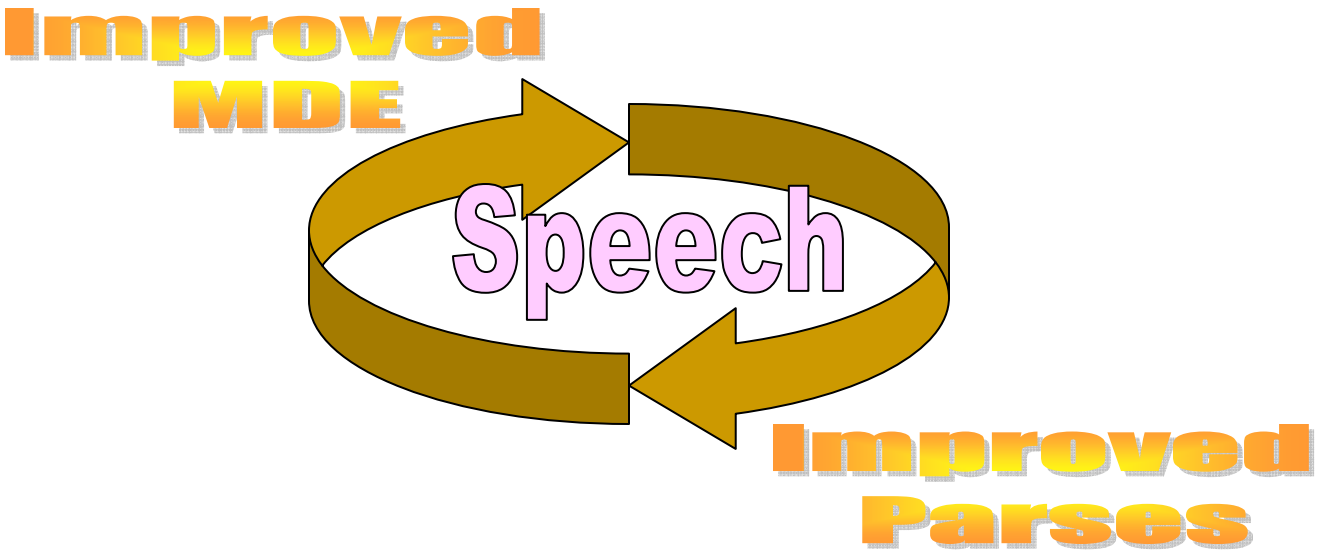
del ins correct

NIST error (per SU event) = 2/2 = 100%

(NIST metric tends to be high)

Per-boundary-based error = 2/6 = 33%

Explore the Synergy: MDE's Impact on Parsing Speech

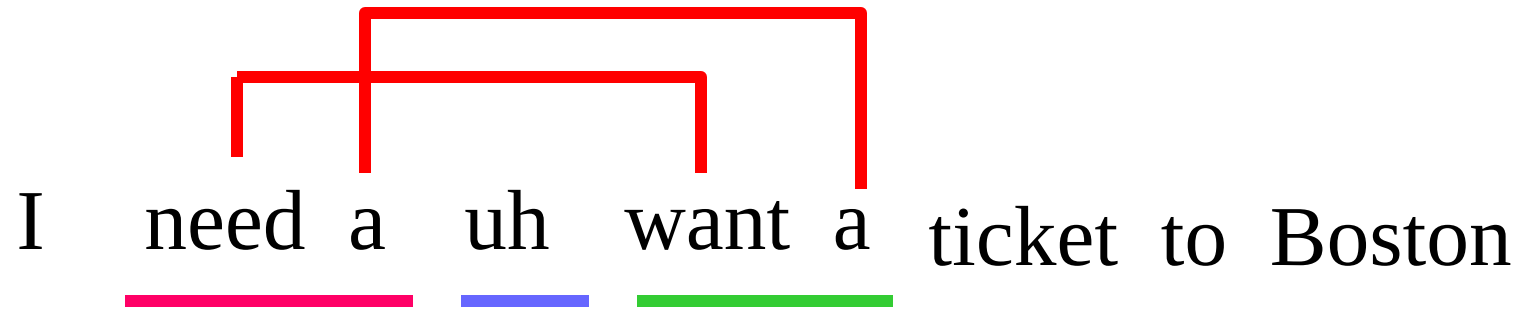


MDE and Parsing

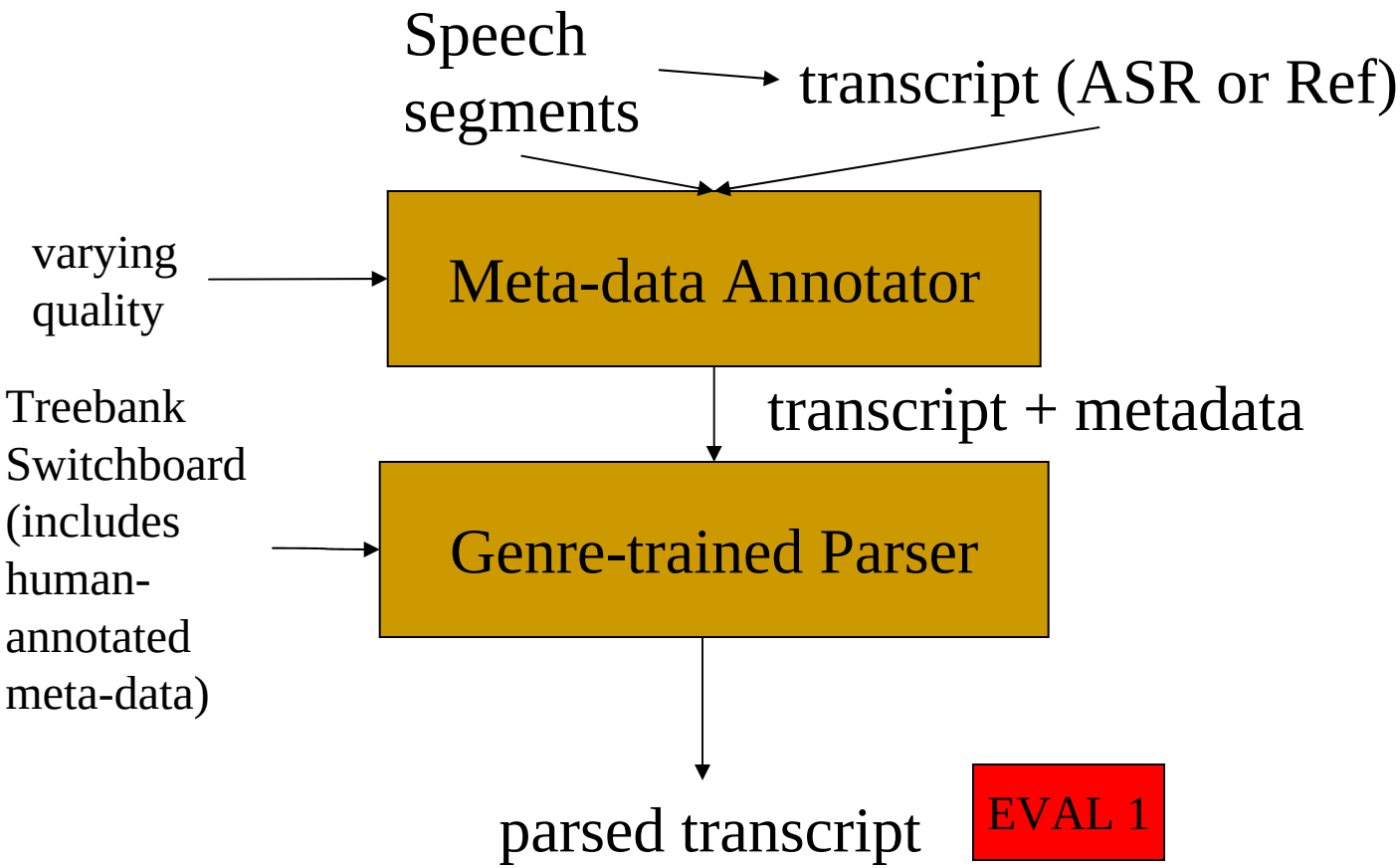
- SU Issues:
  - Could perform supervised training on parsers so that the parser identifies not only the underlying structure of an SU but the underlying structure of a entire conversation side.
  - This approach is infeasible due to issues of computational complexity, not to mention memory issues.
- EDITED Region Issues:
  - Could perform supervised training on parsers so that the parser identifies Edited regions within an SU.
  - Although supervision can be provided via an appropriately annotated treebank and is computationally tractable, there are issues of representational power to address.
- Fortunately, it is a simple matter to pass n-best hypotheses from a structural metadata system to the parser.

Speech Repairs

- Repetitions and Content Replacements speech repairs have a structure that involves cross-serial dependencies.
- Many parsers are unable to model the cross-serial dependencies explicitly (exceptions are TAG and CDG)



Evaluate How MDE Affects Parsing



Measuring Parse Accuracy on Speech

- How do we measure parsing accuracy/error given:
  - Word mismatch
  - SU mismatch
- Alignment issues:
  - reference transcript and ASR output need to be aligned in some manner prior to computing a Parseval type score
  - Work on conversation sides and super-trees
- Metrics: bracket-based (i.e., adapt Parseval metrics), dependency-based, leaf ancestor
- Examples to discuss issues

Evaluation Issues: Word and SU Errors

when i came out of school a lot of the diagnostic procedures done now you have to mix chemicals is yes

```
(S1 (S (SBAR (WHADVP (WRB when))
          (S (NP (PRP i)
             (VP (VBD came)
                 (PP (IN out) (PP (IN of) (NP (NN school))))
                 (S (NP (NP (DT a) (NN lot))
                                (PP (IN of)
                                    (NP (DT the) (JJ diagnostic)
                                        (NNS procedures))))
                    (VP (VBN done))))))
      (ADVP (RB now))
      (NP (PRP you))
      (VP (VBP have) (S (VP (TO to)
                                (VP (VB mix)
                                    (SBAR (S (NP (NNP chemicals))
                                             (VP (VBZ is)
                                             (INTJ (UH yes)))))))))
      (. .)))
```

Issues for Evaluation

- The tree disappears:
  (S1 (XX (X fi-)) (. .)))
  (S1 (EDITED ….))
-

The Gold Standard

when i came out of school a lot of the diagnostic procedures were done manually

```
(S1 (S (SBAR (WHADVP (WRB when))
          (S (NP (PRP i))
            (VP (VBD came)
                 (PP (IN out)(PP (IN of) (NP (NN school)))))))
     (NP (NP (DT a) (NN lot))
              (PP (IN of)
                   (NP (DT the) (JJ diagnostic)
                        (NNS procedures))))
     (VP (VBD were) (VP (VBN done)
                          (ADVP (RB manually))))
     (. .)))
```

had to mix chemicals

```
(S1 (VP (VBD had) (S (VP (TO to) (VP (VB mix)
                                      (NP (NNS chemicals)))))
      (. .)))
```

etc.

Evaluation of Parsed Speech

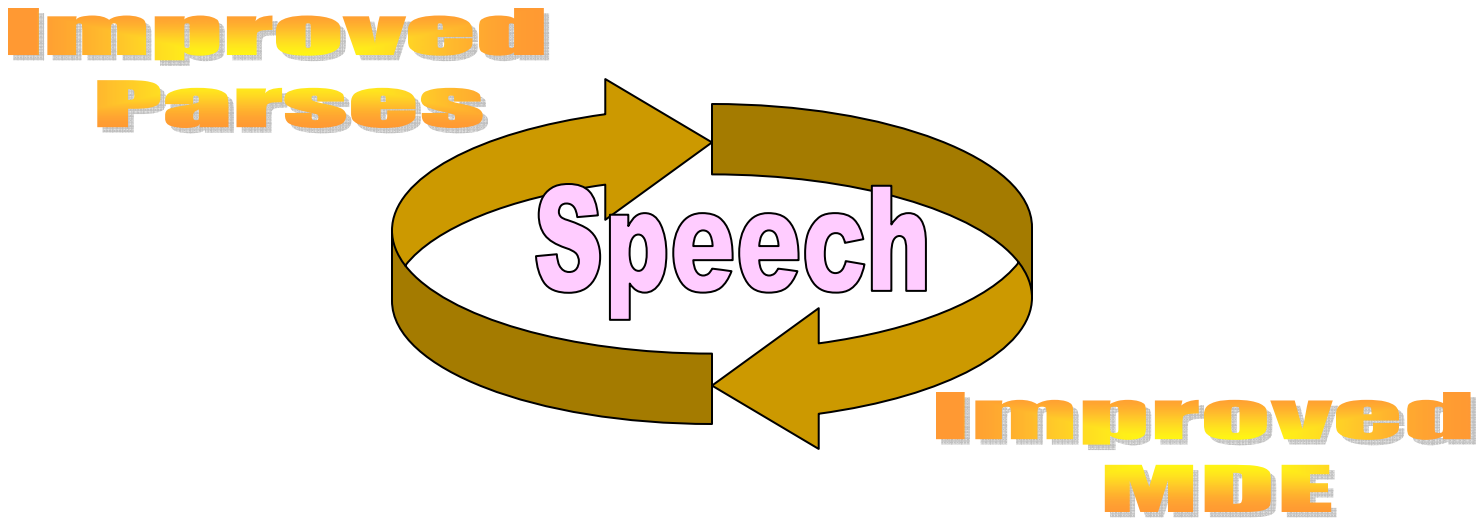
- Need to compare parses over potentially different yields
  - Reference syntactic parse for reference words
  - Automatic syntactic parse for ASR output
- Some metrics rely on externally supplied alignment
  - generalized PARSEVAL metrics (labeled bracketing)
  - Leaf-ancestor
- Head-to-head dependencies can be evaluated either with or without external alignments
- New package (*sparseval*) supports the use of these different metrics
  - also allows open/closed class sensitive evaluation

Impact of Structural Metadata on Parsing **(dev2)**

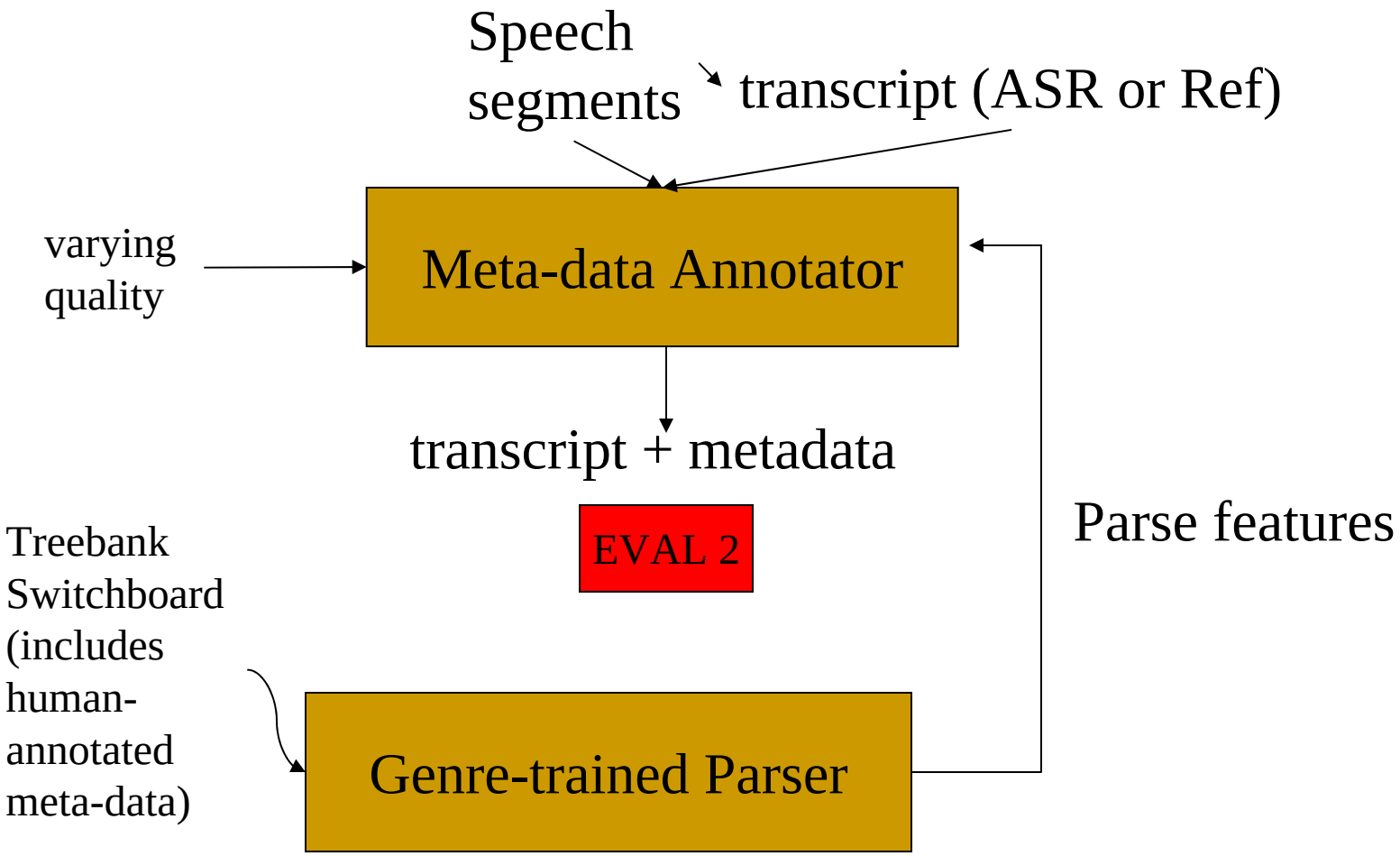
	SU boundary	SU+subtype	edit words
ref words:	**27.30**	**36.89**	**53.39**
stt:	**37.34**	**47.03**	**76.03**

	Human Transcriptions	**ASR Output**
Human-Annotated Metadata	Best for the Parser (84.36 dep F-meas)	Hard to evaluate
System-generated Metadata	The effect of MDE errors (*76.48* dep F-meas)	**The effect of Word and MDE errors (65.24 dep F-meas)**

Explore the Synergy: Syntactic Impact on MDE



Direction 2


Speech segments
transcript (ASR or Ref)
varying quality
Meta-data Annotator
transcript + metadata
EVAL 2
Parse features
Treebank Switchboard (includes human-annotated meta-data)
Genre-trained Parser

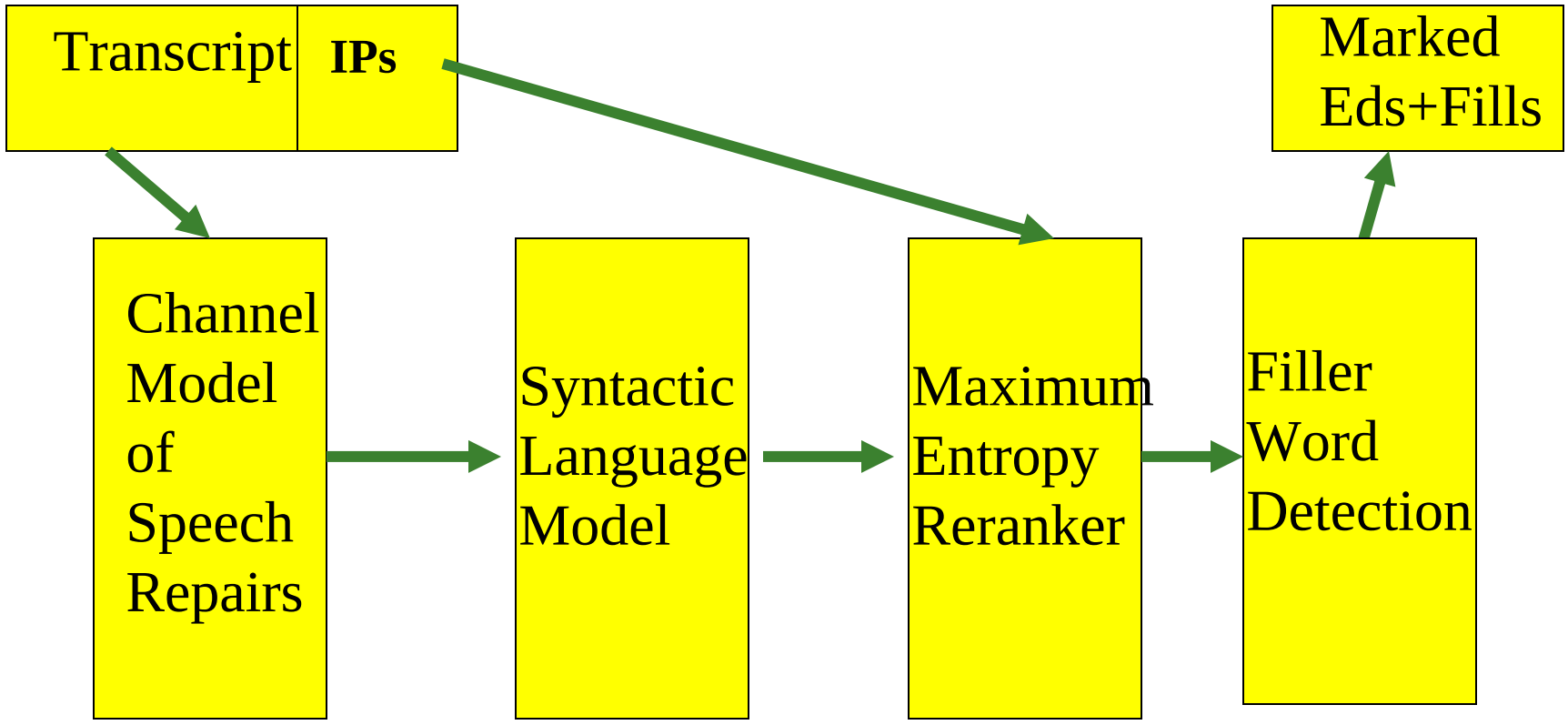
A Noisy-Channel Model for Detecting Repairs

Johnson & Charniak. *A TAG-based noisy-channel model of speech repairs*. ACL 2004.

$$\hat{s} = \arg\max_{S} P(S \mid O) = \arg\max_{S} P(S)P(O \mid S)$$

- Want to recover most likely source sentence (without repair) for a given observed sentence
- P(S): *Language Model* tells us probability of source sentences (first bigram, then parser-based)
- P(O|S): *Channel Model* tells us probability of repair insertions

The System


Transcript
IPs
Marked Eds+Fills
Channel Model of Speech Repairs
Syntactic Language Model
Maximum Entropy Reranker
Filler Word Detection

Results on Evaluation Data

MDE Type	ASR Output	Human Transcript
Edit words	76.2 (80.72)	46.1 (51.49)
Filler words	39.9 (42.53)	23.7 (27.10)
Interruption Point	55.9 (60.59)	28.6 (30.31)

- SUs provided by ICSI/SRI/UW MDE system
- EWD via reranked, noisy-channel model
- FWD via a few simple, deterministic rules
- IPD determined by EWD and FWD predictions

Impact of Parse Information on MDE

	Human Transcriptions	ASR Output
Accurate Parse Information	**Best Case for Metadata Annotation**	Hard to evaluate
Less Accurate Parse Information	How negative of an impact would parse errors have on MDE?	**Worst Case for Metadata Annotation**

Resources

- Corpora and Parse Banks
  - Switchboard Penn Treebank (not entirely consistent with the RT'04 MDE specification).
  - Ears RT'04 metadata corpora
  - New RT'04 Treebank (dev, dev2, eval)
- Various Parsers
- ICSI MDE system
- Other NLP tools: taggers, chunkers, etc.

Value of RT'04 for Evaluation

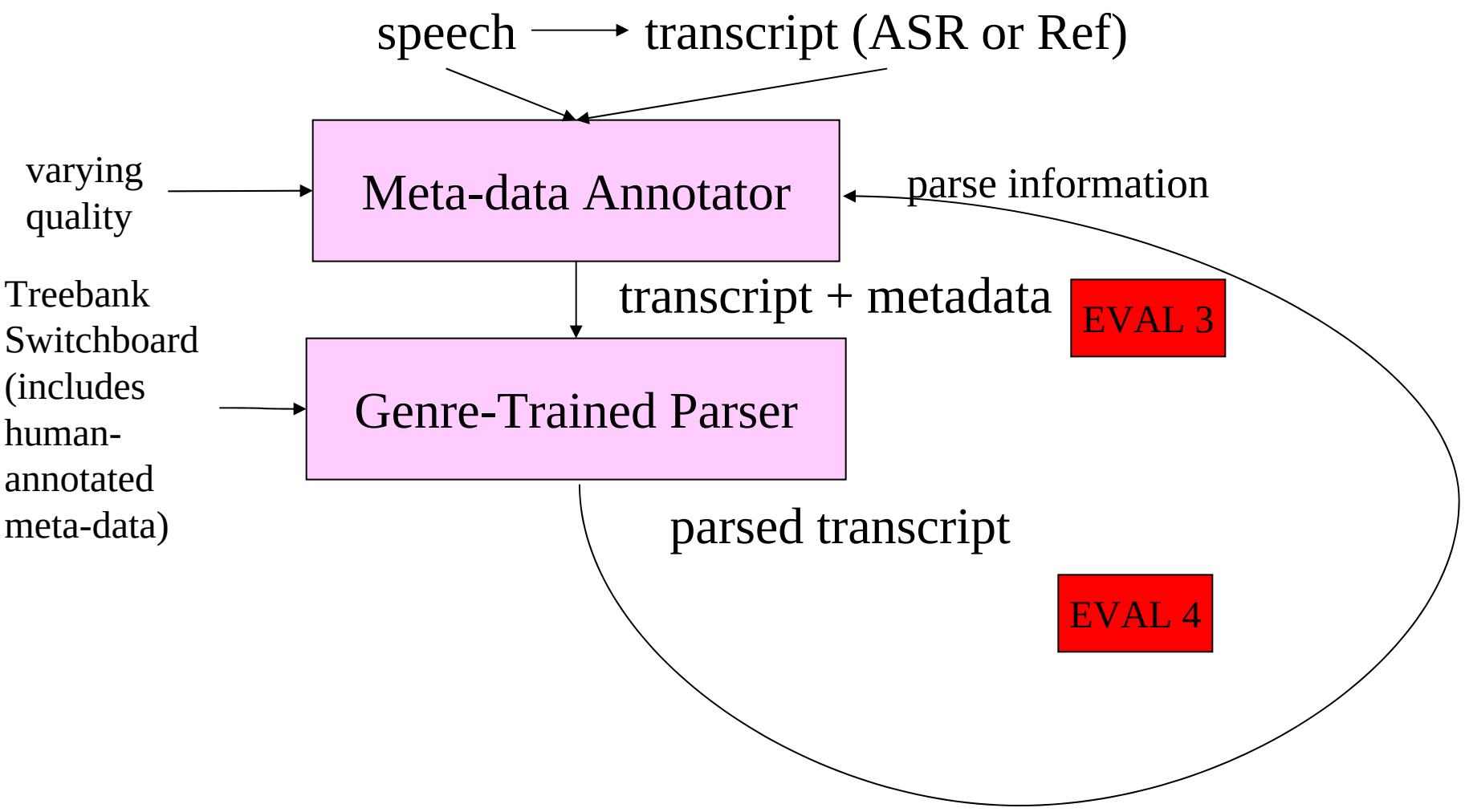
- The RT'04 data is annotated with metadata that has been used in the RT MDE benchmark tests
- There is now gold standard parses from the LDC treebanking team for dev, dev2, and eval sets.
- Recognition output from state-of-the-art recognizers is available for the EARS RT'04 data.
- Using this new data allows us to evaluate the synergy between parsing and MDE system performance.

RT'04	Conversations	#SU	#words
dev	72	11k	71k
dev2	36	5k	35k
eval	36	5k	34k

Pre-Workshop

- Constructed dev, dev2, and eval treebanks for the RT-04 metadata corpora.
- Trained parsers on Switchboard-3 Treebank under various conditions (with and without EDITED).
- Modified metadata system (without parsing) to output n-best MDE hypotheses for reranking.
- Created Sparseval and used it for baselines

Closing the Loop



Acknowledgements

- **The team:** Bonnie Dorr, John Hale, Anna Krasnyanskaya, Matt Lease, Yang Liu, Brian Roark, Zak Shafran, Matt Snover, Robin Stewart, Lisa Yung
- **Others involved:** Eugene Charniak, Mark Johnson, Jeremy Kahn, Mari Ostendorf, Andreas Stolcke, Liz Shriberg, Wen Wang
- **Other assistance:** Fernando Pereira and Andrew McCallum, Ann Bies and the LDC Treebanking Team
- **Our sponsors and everyone at CLSP!**