LINGUISTIC DATA: FAST TRANSCRIPTION

L. Lamel, J.L. Gauvain

RT03 meeting
Boston, MA
May 20, 2003
SIMPLIFYING TRANSCRIPTION TASK

- 20-40xRT manual transcription time is severely limits the quantity of data
- Much effort spent on difficult portions that are not used for training
  - verifying spelling of proper names
  - transcribing overlapping speech
  - precision annotation of speaker turns and acoustic conditions
- Better to have more data with less detailed transcriptions
METHOD USED FOR DEV03

- Transcribe data with a good BN STT system
- Should language model include captions?
- Should we use source-specific language models?
- Align hypotheses with captions (when available), highlighting errors
01.35-05.34 from public radio international this is the world
07.00-12.25 OF (A) co production of the b. b. C’S (C.) world service p. r. i. and w. g. b. h. boston
17.01-20.52 it’s wednesday january seventeenth i’m lisa mullins in boston
25.07-29.45 today colin powell SERVED AS (SURVEYS) the world’s HOTSPOTS (HOT SPOTS) including the persian gulf
29.45-37.04 WITH A (WHEN WE) look at that whole troubled region mr. chairman there’s no more tragic case THAT (THAN) iraq THE VEIL (A FAILED) state with a failed leader

Manually corrected output:
01.35-05.34 from public radio international this is the world
07.00-12.25 a co production of the b. b. c. world service p. r. i. and w. g. b. h. boston
17.01-20.52 it’s wednesday january seventeenth i’m lisa mullins in boston
25.07-29.45 today colin powell surveys the world’s hot spots including the persian gulf
29.45-37.04 when we look at that whole troubled region mister chairman there is no more tragic case than iraq a failed state with a failed leader
EXAMPLE CORRECTION RULES

Caption Correct

IN (AND) ⇒ and  
AND (IN) ⇒ in  
ALIVE (A LIVE) ⇒ a live  
BUILD (BUILT) ⇒ built  
FIFTEEN (FIFTY) ⇒ fifty  
FOR (FOUR) ⇒ four  
TO (TWO) ⇒ two  

Recognizer Correct

DOT (POINT) ⇒ dot  
DOCTOR (DR) ⇒ doctor  
A (ONE) ⇒ a  
THE SECOND (TWO) ⇒ the second  
million (DOLLARS) ⇒ million  
billion (DOLLARS) ⇒ billion  
thousand (DOLLARS) ⇒ thousand  

About 120 correction rules
MANUAL POST PROCESSING

- Verification and marking of commercials (removed from scoring)
- Verification of some proper names
- Replace \{FW\} by (\%hesitation), remove \{breath\}
- Expansion of contractions by context
  - it’s → it has / it is
  - he’d → he had / he would
  - we’re → we are / we were

Estimated total correction time: \(\sim 5 \times RT\)
CONCLUSIONS

• Proposed method to speed up transcription for acoustic and language model training

• Need comparative experiments to determine best choice for language model

• Need comparative experiments to more accurately assess time for manual correction

• Fast transcripts be further refined for detailed manual transcripts of test data