Using OntoNotes PropBank and Syntax

Dilek Hakkani-Tür

dilek@icsi.berkeley.edu

Collaborators: Benoit Favre, ICSI

Sibel Yaman, ICSI

Gokhan Tur, SRI

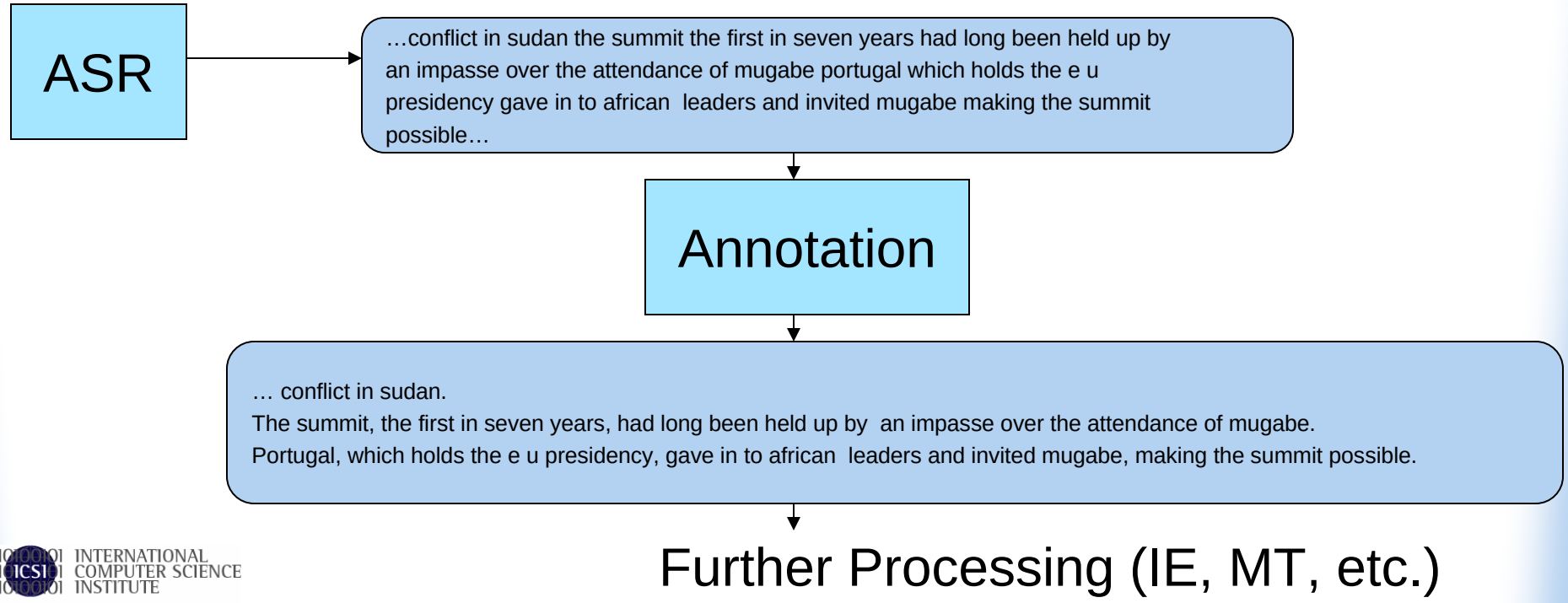
INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Outline

- Syntax for Punctuation Insertion
  - Sentence Boundaries
  - Comma Boundaries
- Syntax and Semantics for Extraction of 5Ws
- Feedback/Conclusions

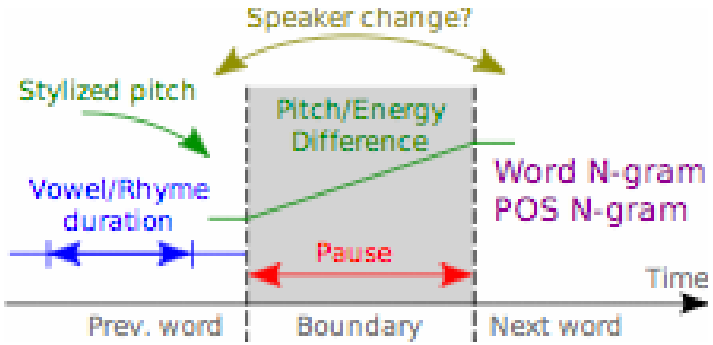
Punctuation Insertion for Speech

- Speech Recognizer Output:
  - Lacks punctuation and capitalization.
  - Includes transcription errors
- Sentences and commas make reading easier for humans, but also processing easier for machines.



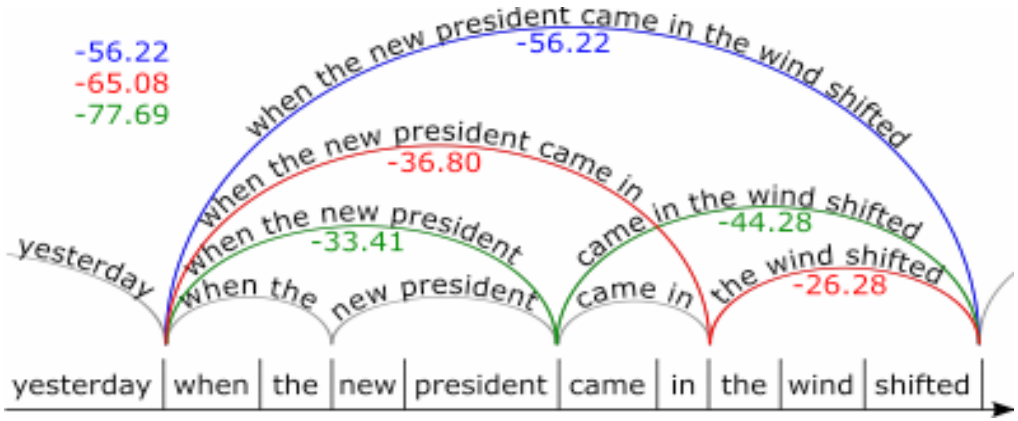
Sentence Segmentation

- Word boundary classification with local features.

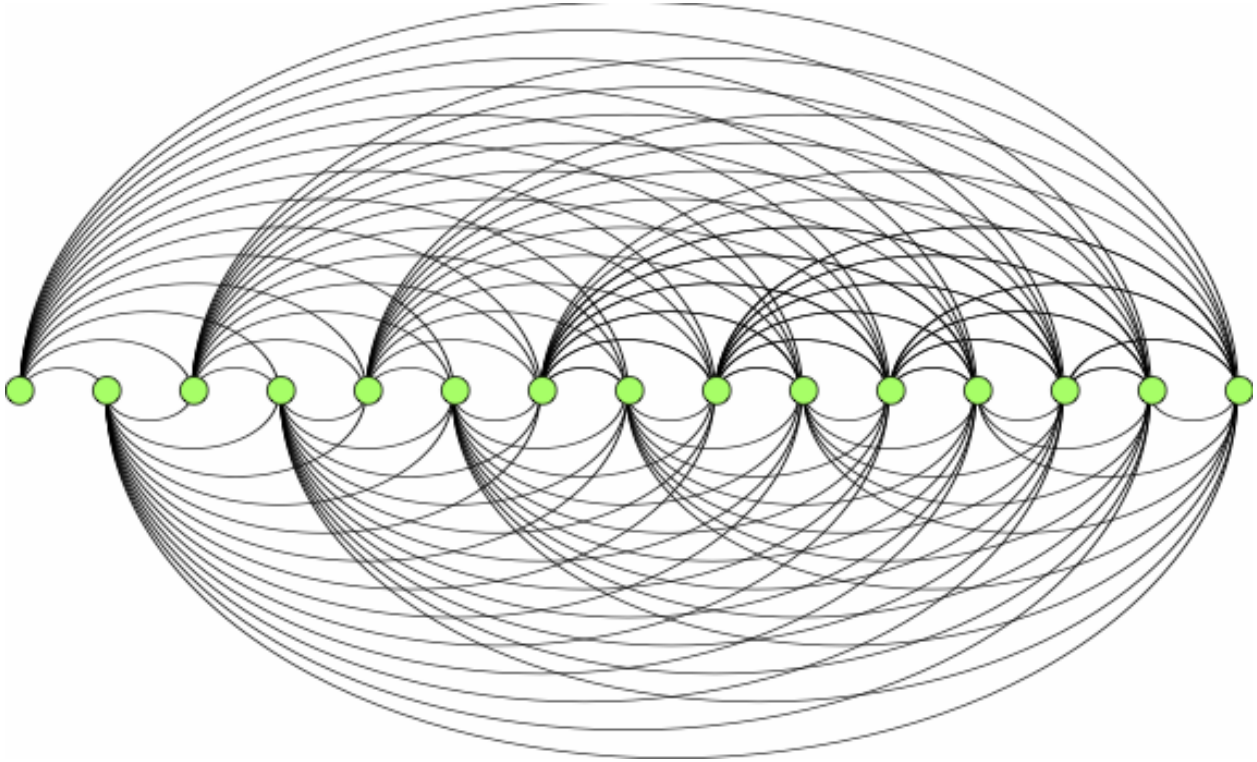


- Features:
  - Lexical: words
  - Structural: speaker changes
  - Prosodic: Pitch, energy, and durations
- Local and sequence classifiers.

- Using syntax: use parser as a language model?

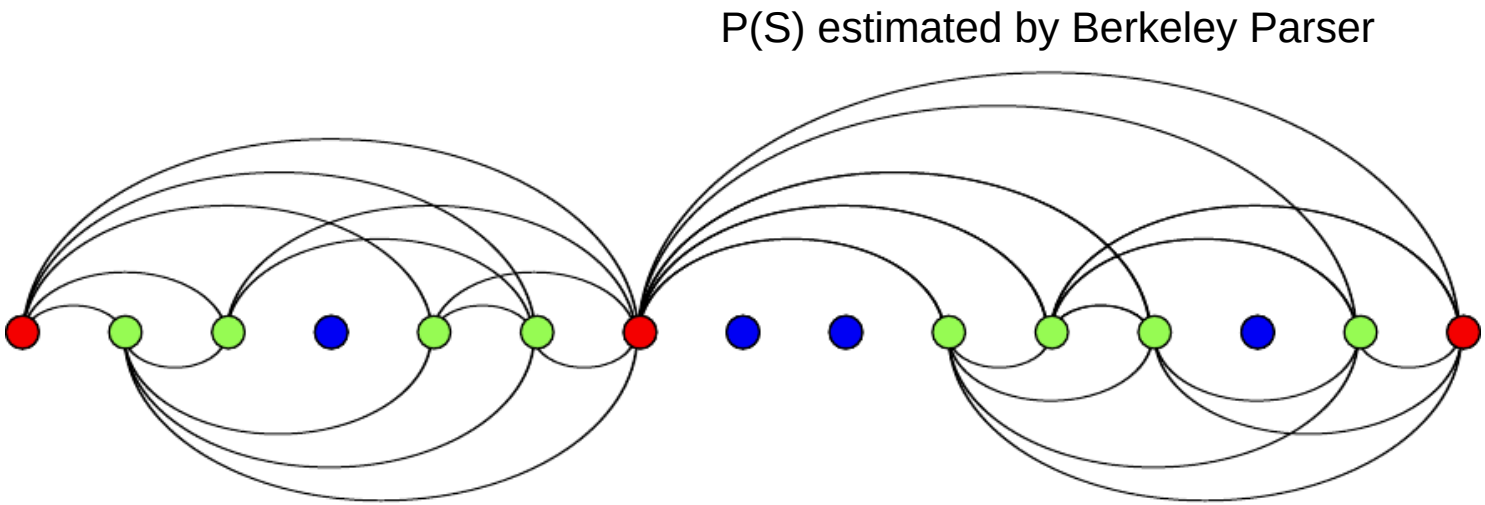


Forming the Sentence Hypothesis Lattice



- Huge search space
- Slow parsers

Forming the Sentence Hypothesis Lattice



- Non-sentence boundary with a high score according to the local model
- Sentence boundary with a high score according to the local model
- Candidate boundary

Experiments

- F-measure on TDT-4 test set
- Local Classifier: Boosting

	Mandarin (F-measure)
Boosting	68.8
Boosting + syntax (BP, WSJ)	68.7
Boosting + syntax (BP, TDT-4 ASR, self-training)	70.1
Boosting + syntax (BP, OntoNotes)	71.1

- Training parser for ASR and training parser for speech both help.
- Future challenges:
  - supervised training using ASR data with reference parse trees.
  - Parsing for segmenting conversational speech

Conclusions and Feedback

- Training parsers for genre is beneficial.
  - More speech data (both conversations and news).
- Richer annotation for speech data:
  - Disfluencies including types (repair, repeat, filled pause, etc.).
  - Dialog act tags (e.g. backchannels and interruptions for conversations)

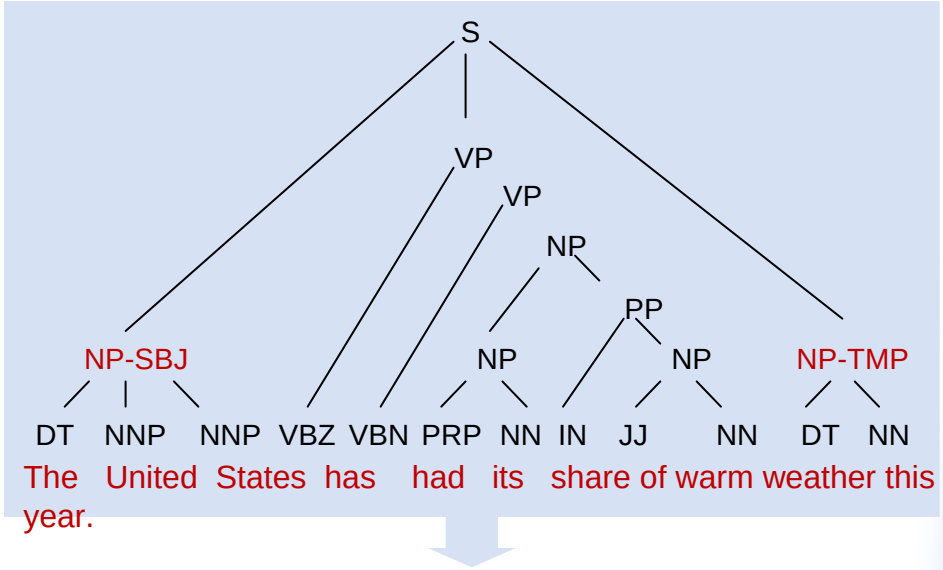
Outline

- Syntax for Punctuation Insertion
  - Sentence Boundaries
  - Comma Boundaries
- Syntax and Semantics for Extraction of 5Ws
- Feedback/Conclusions

Answer Extraction for 5W Questions

- Given 5W questions, find the exact answer.
- Find the 5Ws of sentences.

- Syntactic parses with function tags
  - UMD parser (M. Harper)
  - SRI parser (W. Wang)
  - Berkeley parser
- Preprocessing:
  - Detect quotes
  - Find passives
  - Find top level predicates
- Rules



WHO: The United States
WHAT: has had its share of warm weather
WHERE: null
WHEN: this year
WHY: null

Answer Extraction for 5W Questions

- Semantic role labeler
  - Based on Pradhan et al.'s ASSERT.
  - Trained using OntoNotes PropBank annotations.
  - Uses Berkeley Parser with function tags, also trained using OntoNotes.

TARGET: **had**
ARG0: The United States
ARG1: its share of warm weather
ARGM-TMP: this year

WHO: The United States
WHAT: has had its share of warm weather
WHERE: null
WHEN: this year
WHY: null

Experiments for 5W Extraction

- Re-scoring/Re-ranking to select one output.
  - 2-state and 6-state HMM.
- Experiments with OntoNotes reference transcriptions, split into train/dev/test:

WHO	Parser Correct	Parser Incorrect
SRL Correct	90.8%	2.9%
SRL Incorrect	4.9%	1.4%

Best (Parser): 95.7%
Oracle: 98.6%
Merged: 94.0%

WHAT	Parser Correct	Parser Incorrect
SRL Correct	61.4%	11.9%
SRL Incorrect	11.3%	15.4%

Best (SRL): 73.3%
Oracle: 84.6%
Merged: 74.8%

WHERE	Parser Correct	Parser Incorrect
SRL Correct	39.7%	12.5%
SRL Incorrect	20.1%	27.7%

Best (Parser): 59.8%
Oracle: 72.3%
Merged: 67%

Conclusions and Feedback

- Training parsers for genre is beneficial.
  - More speech and web data.
- Function tags help 5W answer extraction, even though OBJ is not labeled, and hence wasn't generated by the parser.
- Data from older dates: is ASR using them already?
- Active Learning:
  - Select data for parsing, for SRL, for ASR?