Combining Words and Prosody for Information Extraction from Speech

Andreas Stolcke

Elizabeth Shriberg

Dilek Hakkani-Tür

Gökhan Tür

Ze'ev Rivlin

Kemal Sönmez

SRI International

Beyond Speech Recognition

- Speech Understanding is our long-term goal
- Short-term: learn to extract useful structures and elements of meaning from speech, including
  -sentence boundaries
  -topic boundaries (for topic tracking/detection)
  -named entity (NE) recognition
- Most current techniques are text-based --- but speech is missing important cues (punctuation, capitalization, paragraphs, headers, etc.)
- We are not using information specific to speech

Prosody for Information Extraction

- Idea: pitch and duration of speech units contain important cues
  - for segmentation (sentences, topic)
  - about what's NEW and IMPORTANT (possibly helpful to find NEs)

- Research Issues:
  - How can prosodic information be extracted?
  - How can cues be modeled ?
  - How to combine them with word-based cues?
  - Do they help on our tasks?

Overview of Talk

- Topic Segmentation
  - Modeling
  - Results
  - Features
- Named Entity recognition
  - Modeling
  - Results
  - Analysis
- Future Directions
- Conclusions

Topic Segmentation

- Task: Find topic boundaries in BN shows
- Word-based model similar to Dragon HMM
  - states correspond to topic clusters
  - states emit sentences using unigram likelihoods
  - optimized topic transition penalties
  - Viterbi algorithm find best segmentation
- SRI Improvements:
  - added states for (optional) topic-initial and topic-final sentences (5% relative error reduction)
  - topic transitions have additional likelihoods derived from prosody at sentence boundary

Pseudo-Sentence Chopping

- Topic-LMs needs sentence-length units.
  How to pre-segment non-written language?
- Results using word LMs on correct words:

Chop at	Error %
Every 15th word	21.75
Turns	22.50
Sentences	20.72
Pauses > 650 ms	20.06

- Prosodic criterion (pause) works best
- Chopping parameters optimized on held-out set

Prosodic Modeling

- Speakers mark topic boundaries prosodically.
- Decision trees estimate *P*(boundary|prosody)
- Training data downsampled
  - to provide sensitivity to infrequent classes
  - to make posteriors proportional to likelihoods
- Feature pool: pause, duration, and pitch, numerous normalizations and derived features
- Feature subset selection chooses good input set for DT using heuristics and brute force search

Results

- Topic LMs trained on BN'96 and TDT-2 corpus; Prosody models trained on BN'97 acoustic data subset (100x less data than LM training)
- Test on BN'97 subset (comparable toTDT-2 eval)

Model	True Words	Rec. Words
LM only	20.06	20.70
Prosody only	16.51	17.82
Combined	15.01	15.67

- Prosody alone is better than words, and combined model gives substantial additional win (25% rel.)

Feature Usage

- Pitch range and contour features (45.2%)
  - based on stylized, parameterized F0 model
  - 2/3 from range/contour of last word
  - 1/3 difference across boundary
- Pause duration (31.2%)
- Word count (position relative to start) (9.4%)
- Speaker change (8.1%)
- Phone duration (3.8%)

Prosody and Named Entities

- Speakers mark important words prosodically
- Question: What is the correlation between NEs and what speakers consider important?

Model

- Based on HMM name tagger [BBN]
- Decision tree likelihoods attached to HMM states
- Prosody model distinguishes NE, non-NE only

Results

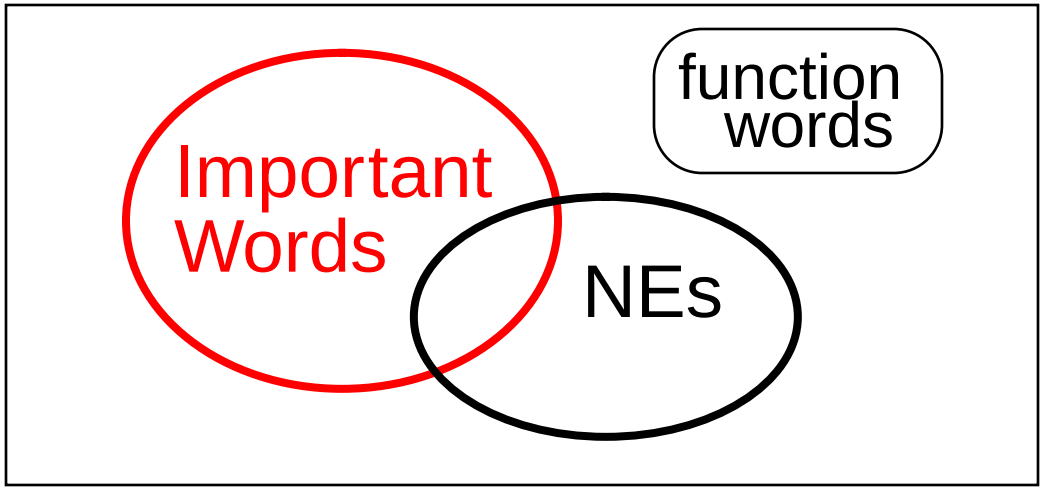
- Prosody alone distinguishes NE/non-NE on an equal priors testset with 69.3% accuracy.
- But: no win from combining prosodic likelihoods with word-based HMM.

What's going on?

- NEs are not always prosodically prominent (e.g., *President Clinton ... Mr. Clinton*)
- Non-NEs can be prominent when denoting new or focussed information (e.g., *earthquake*, *bomb*...)

Diagnostic Experiments

- Experiment shows win from prosody model disappears if we remove function words
- Analysis of prosodically labeled broadcast corpus shows non-NEs more often prominent than NEs (Mari Ostendorf et al., BU)



Future Directions

- Topic segmentation
  - use other prosodic features for chopping
  - integrate sentence and topic segmentation
  - explore alternative classifiers
- Named Entities
  - predict which NEs are not prominent (given)
  - define alternative task for “important” words
- Explore prosody for other tasks ?

Conclusions

- Prosody is an untapped knowledge source for information extraction from speech
- Decision trees are effective prosodic models that can be combined with word-based HMMs
- For topic segmentation, prosody is as good or better than word-based models alone; combined model is even better (25% relative error reduction)
- For NE recognition, prosody gives no win over words. Analysis suggests only partial overlap between NEs and information-bearing words.