Update on
Rich Transcription Surface Form (RT-S)

Matthew Snover (UMD)
Richard Schwartz (BBN)
Bonnie Dorr (UMD)
John Makhoul (BBN)
Overview

- RT-S Review
- Previous Experiments and Results
- New HMM Method
- English CTS Results
- English BNews Results
- Capitalization Scoring
RT-S Task Goal: Produce a complete and readable transcript of speech using normal written conventions.
- Create output that looks like normal transcripts

R: So, have you made any changes?
L: Uh, not particularly.
R: Not particularly? [NOISE]
L: I mean, I, uh -- I’m a …
R: [SIGH]
L: … volunteer in an ambulance squad so it’s affected the squad quite a bit but --
R: Has it affect your [NOISE] activities on the squad?
L: No, just the way we run things.
Previous Experiments

- Five punctuation types: commas, periods, question marks, exclamation points, and discontinuities
- Tested on Fisher Wordwave Eval03 (annotated 5 times)
- Developed MaxEnt automatic punctuation identification system
- Developed scoring metric
  - \( \text{PER} = \frac{\text{# punctuation errors}}{\text{# punctuation in ref}} \)
  - Simple modification to SCLITE gives joint alignment of punctuation and words
  - Cost of punctuation error is 100 times less than word error
  - Scoring with multiple references: hypothesis is correct if it matches any reference
## Previous Results and Conclusions

<table>
<thead>
<tr>
<th></th>
<th>Inter-Annontator</th>
<th>Auto. System Diff Ref Words</th>
<th>Auto. System Same Ref Words</th>
<th>Auto. System w/ ASR Words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Annotator</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PER - WER</td>
<td>31.3</td>
<td>45.8</td>
<td>48.6</td>
<td>47.9</td>
</tr>
<tr>
<td></td>
<td>(10.5 WER)</td>
<td>(10.5 WER)</td>
<td>(0.0 WER)</td>
<td>(25.4 WER)</td>
</tr>
<tr>
<td><strong>4 Annotators</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>11.4</td>
<td>19.1</td>
<td>NA</td>
<td>32.3</td>
</tr>
<tr>
<td>PER - WER</td>
<td>7.6</td>
<td>15.3</td>
<td></td>
<td>13.5</td>
</tr>
<tr>
<td></td>
<td>(3.8 WER)</td>
<td>(3.8 WER)</td>
<td></td>
<td>(18.8 WER)</td>
</tr>
</tbody>
</table>

- High interannotator disagreement; decreases with 4 references
- Automatic system has 8-10% higher error
- PER increases directly with WER
MaxEnt vs. HMM Model
Maximum Entropy Model

- Input is a word transcription
- Output is a punctuated sequence of words
- Treat the task as Part-Of-Speech tagging with punctuation instead of parts of speech
  - Each word is labelled with the punctuation that comes after it, or with the NULL label
  - We trained a maximum entropy tagger (Ratnaparkhi 1996) on 2 million words (~200 hours of speech)
  - Uses 2 words on each side as features
- Maximizes:

\[
Pr(\text{context, tag}) = \pi \mu \prod_{j=1}^{k} \alpha_j f_j(\text{context, tag})
\]
- Hard to train with a lot of data (does not scale well)
HMM Model for Punctuation

- Model punctuation as hidden states
- Uses SRI-LM Toolkit
- Find sequence of punctuation (p) that maximizes:

\[
\prod_{i=1}^{m} \Pr(w_i | p_{i-1} w_{i-1} p_{i-2} w_{i-2} p_{i-3} w_{i-3})
\]

- Bigram version of HMM:

\[
\Pr(w_i \emptyset | w_{i-1} . )
\]

This path means there is a period after \( w_{i-1} \) and no punctuation after \( w_i \).
English CTS Results

- Three punctuation types: end of sentence (EOS), commas, discontinuities

<table>
<thead>
<tr>
<th></th>
<th>Diff Ref Words (10.5 WER)</th>
<th>ASR (25.4 WER)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxEnt (200 hrs)</td>
<td>52.2</td>
<td>70.2</td>
</tr>
<tr>
<td>HMM (200 hrs)</td>
<td>50.8</td>
<td>68.7</td>
</tr>
<tr>
<td>HMM (1500 hrs)</td>
<td>49.7</td>
<td>67.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 References</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxEnt(200 hrs)</td>
<td>17.8</td>
<td>31.3</td>
</tr>
<tr>
<td>HMM (200 hrs)</td>
<td>16.7</td>
<td>30.6</td>
</tr>
<tr>
<td>HMM (1500 hrs)</td>
<td>15.9</td>
<td>29.9</td>
</tr>
</tbody>
</table>

- HMMs slightly outperform MaxEnt model
- More data doesn’t help very much
English CTS Results By Type

- Results with different reference words

<table>
<thead>
<tr>
<th></th>
<th>Comma</th>
<th>Discont</th>
<th>EOS</th>
<th>Total PER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Reference</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxEnt (200 hrs)</td>
<td>50.5</td>
<td>83.4</td>
<td>36.4</td>
<td>52.5</td>
</tr>
<tr>
<td>HMM (200 hrs)</td>
<td>53.7</td>
<td>68.4</td>
<td>35.4</td>
<td>50.8</td>
</tr>
<tr>
<td>HMM (1500 hrs)</td>
<td>52.5</td>
<td>66.0</td>
<td>35.1</td>
<td>49.7</td>
</tr>
<tr>
<td><strong>4 References</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MaxEnt (200 hrs)</td>
<td>12.9</td>
<td>30.9</td>
<td>18.9</td>
<td>17.8</td>
</tr>
<tr>
<td>HMM (200 hrs)</td>
<td>13.4</td>
<td>21.7</td>
<td>17.8</td>
<td>16.7</td>
</tr>
<tr>
<td>HMM (1500 hrs)</td>
<td>12.7</td>
<td>20.6</td>
<td>17.3</td>
<td>15.9</td>
</tr>
</tbody>
</table>

- HMM is much better at discontinuity detection
- Equivalent performance on commas and sentence boundaries
- Commas around filled pauses are ‘free’
- Many EOS are at end of speaker turns
English Broadcast News

- Differences from CTS:
  - Only end-of-sentence markers and commas
  - Longer speaker turns
- HMM trained on TDT3 and English Gigaword Corpus
  - equivalent to over 100,00 hrs
  - HUB4 is BNews data, but has no commas
  - Gigaword is text data
- Tested on TDT4 Data

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Comma</th>
<th>EOS</th>
<th>Total PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDT3, Gigaword</td>
<td>82.1</td>
<td>74.8</td>
<td>78.2</td>
</tr>
<tr>
<td>TDT3, Gigaword, Hub4</td>
<td>95.9</td>
<td>55.5</td>
<td>74.2</td>
</tr>
</tbody>
</table>

- BNews is much better than text in Gigaword Corpus for EOS
- Much higher error rate than CTS
Capitalization Error Rate
Measuring Capitalization Error Rate (CapER)

- Use similar techniques to measure CapER
  - CapER = \(<\# \text{ capitalization errors}\> / \(<\# \text{ capitalized ref words}\>\)
  - If a word is substituted, but capitalization is the same, then no capitalization error
  - If a capitalized word is deleted, it is a capitalization error

- SCLITE performs joint alignment of words, punctuation, and capitalization
  - Cost of cap. error < cost of punc. error < cost of word error

Ref: Well she saw The Nightmare Before Christmas ***
Hyp: He *** saw *** nightmare before Christmas Day
Eval: D D D D I
Capitalization Model and Results

- Simple unigram model. If a word (in a non-sentence initial position) is capitalized more than 50% of time in training data, then capitalize it in test data.
- Always capitalize first word of the sentence.
- More scientifically interesting is error rate on non-sentence initial words: CapER*
  - Capitalization at start of reference or hypothesis sentences are ignored.
  - Measures capitalization of proper nouns (in English).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CapER</th>
<th>CapER*</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS Ref Punc</td>
<td>9.6</td>
<td>19.5</td>
</tr>
<tr>
<td>Auto Punc</td>
<td>23.0</td>
<td>16.7</td>
</tr>
</tbody>
</table>
Summary

- CTS: HMM punctuation model has lower error rate for discontinuities than MaxEnt model
- HMMs are ‘easier’ to use with more data, but more data doesn’t seem to help very much for CTS
- High error rate for BNews
  - Longer speaker turns
  - A lot of punctuation is ‘free’ in CTS
- Proposed capitalization scoring metric (CaPER)
What’s Next?

• Measure the effect of automatic punctuation on Information Extraction
• Add simple prosodic features
• Build separate model to distinguish periods, question marks and exclamation points
  – Use prosodic features and presence of ‘question’ words
• Expand to Mandarin and Arabic