Transcriber driving strategies for transcription aid system

Grégory Senay, Georges Linarès, Benjamin Lecouteux, Stanislas Oger

Laboratoire Informatique d’Avignon

LREC’2010 - May 2010
Overview

- Introduction
- What is interactive decoding?
- Driving strategies
- Experiences and results
- Conclusion
Introduction

Current situation

- Automatic Speech Recognition system performance:
  - ⇒ accurate on defined domains (ex: Broadcast news)
  - ⇒ decreases, if the conditions are changed
- Manual transcriptions are needed to provide a perfect transcription
- Recent projects use transcriptions provided by a speech recognition system
  - ⇒ they only use the one-best hypothesis [Bazillon LREC08]

Objective

- Reduce the cost of the global transcription
- Correction efficiency
- Computer and Human can work together
Interactive decoding

Description

- It is a semi-automatic transcription task, in 2 steps:
  - human correction
  - a fast decoding pass
- ASR system evaluates a lot of alternatives paths
- Different alternatives could be proposed to the transcriber
- We use Confusion Network: more readable than lattice
Interactive decoding

⇒ First pass and CN generation

ASR

CONFUSION NETWORKS

W1 W2 W3 W4 W5

LREC’2010 - Grégory Senay
Transcriber driving strategies for transcription aid system
Interactive decoding

⇒ Transcriber makes a correction
The correction is integrated into the re-decoding step
⇒ A new confusion network is generated with a new transcription

⇒ CN is reduced
Interactive decoding with driving strategies

⇒ Methods drive the transcriber to the critical areas

LREC’2010 - Grégory Senay
Transcriber driving strategies for transcription aid system
Driving Strategies

Left-Right

- In the reading direction
- A normal strategy for the transcriber
- Drives on the left to the right

Graph density

- Numerous methods use graph density as a confidence measure
- The deepest part of a graph is a critical area where system has trouble to choose between a large number of hypotheses
- Graph density drives toward the widest section of the Confusion Network
Driving Strategies - Semantic consistency

2 methods are used: based on Corpus and Web

⇒ Each segment is split in small windows (10 relevant words)
⇒ The transcriber is driven to the lowest score window

Corpus criterion
- Principle: find in the corpus the closest newswire
- Based on a large corpus of newswires: Gigaword
  - 2 millions of newswires - 250 millions of sentences
- Corpus score is performed by the Cosine metric

Web criterion
- Web has a large language coverage
- Each Web documents is regarded as a bag-of-words
- Web score: words co-occurrence probability on the Web
Experiments - Protocol

Broadcast news system

- LIA broadcast news system: SPEERAL
- Development framework of the ESTER campaign
  - 8 hours from 4 different radio stations
- System on first pass: 32.6% Word Error Rate
  - 2 x Real Time
  - without speaker adaptation
  - first pass produces Confusion Networks
- Transcription is automatically split according to:
  - speaker turns
  - silence areas
  - length (30 seconds maximum)
**Interactivity**

- Corrections are simulated by *Sclite*
- \[ WER = \frac{\text{confusion} + \text{insertion} + \text{deletion}}{\# \text{word number}} \]
- Re-decoding on Real Time system

**Results**

- Corrections start from the ASR transcriptions
- The baseline: **Human only** (without interactive decoding)
- Global WER evaluated for each correction
- 2 classes: below and above 40% WER
WER of corrections for initial transcriptions of WER below 40%.

<table>
<thead>
<tr>
<th># c/segment</th>
<th>1</th>
<th>3</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human only</td>
<td>25.22</td>
<td>22.98</td>
<td>17.23</td>
<td>9.44</td>
</tr>
<tr>
<td>LR-ID</td>
<td>24.28</td>
<td>20.82</td>
<td>11.88</td>
<td>5.26</td>
</tr>
<tr>
<td>GD-ID</td>
<td>26.58</td>
<td>25.38</td>
<td>16.62</td>
<td>11.76</td>
</tr>
<tr>
<td>Corp-ID</td>
<td>23.90</td>
<td>21.15</td>
<td>13.93</td>
<td>8.51</td>
</tr>
<tr>
<td>Web-ID</td>
<td>24.33</td>
<td>21.10</td>
<td>12.21</td>
<td>7.40</td>
</tr>
</tbody>
</table>

ID: Interactive Decoding
WER of corrections for initial transcriptions of WER below 40%.

<table>
<thead>
<tr>
<th># c/segment</th>
<th>1</th>
<th>3</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human only</td>
<td>25.22</td>
<td>22.98</td>
<td>17.23</td>
<td>9.44</td>
</tr>
<tr>
<td>LR-ID</td>
<td>24.28</td>
<td><strong>20.82</strong></td>
<td><strong>11.88</strong></td>
<td><strong>5.26</strong></td>
</tr>
<tr>
<td>GD-ID</td>
<td>26.58</td>
<td>25.38</td>
<td>16.62</td>
<td>11.76</td>
</tr>
<tr>
<td>Corp-ID</td>
<td><strong>23.90</strong></td>
<td>21.15</td>
<td>13.93</td>
<td>8.51</td>
</tr>
<tr>
<td>Web-ID</td>
<td>24.33</td>
<td>21.10</td>
<td>12.21</td>
<td>7.40</td>
</tr>
</tbody>
</table>

ID: Interactive Decoding
WER of corrections for initial transcriptions of WER below 40%.

<table>
<thead>
<tr>
<th># c/segment</th>
<th>1</th>
<th>3</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human only</td>
<td>25.22</td>
<td>22.98</td>
<td>17.23</td>
<td>9.44</td>
</tr>
<tr>
<td>LR-ID</td>
<td>24.28</td>
<td>20.82</td>
<td>11.88</td>
<td>5.26</td>
</tr>
<tr>
<td>GD-ID</td>
<td>26.58</td>
<td>25.38</td>
<td>16.62</td>
<td>11.76</td>
</tr>
<tr>
<td>Corp-ID</td>
<td>23.90</td>
<td>21.15</td>
<td>13.93</td>
<td>8.51</td>
</tr>
<tr>
<td>Web-ID</td>
<td>24.33</td>
<td>21.10</td>
<td>12.21</td>
<td>7.40</td>
</tr>
</tbody>
</table>

ID: Interactive Decoding
WER of corrections for initial transcriptions of WER above 40%.

<table>
<thead>
<tr>
<th># c/segment</th>
<th>1</th>
<th>3</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human only</td>
<td>55.91</td>
<td>54.05</td>
<td>47.81</td>
<td>40.14</td>
</tr>
<tr>
<td>LR-ID</td>
<td>54.95</td>
<td>49.77</td>
<td>37.71</td>
<td><strong>25.36</strong></td>
</tr>
<tr>
<td>GD-ID</td>
<td>57.51</td>
<td>53.52</td>
<td>44.05</td>
<td>36.99</td>
</tr>
<tr>
<td>Corp-ID</td>
<td>54.19</td>
<td>49.37</td>
<td>39.06</td>
<td>29.54</td>
</tr>
<tr>
<td>Web-ID</td>
<td><strong>51.88</strong></td>
<td><strong>48.32</strong></td>
<td><strong>37.49</strong></td>
<td>29.49</td>
</tr>
</tbody>
</table>

ID: Interactive Decoding
WER of corrections for initial transcriptions of WER above 40%.

<table>
<thead>
<tr>
<th># c/segment</th>
<th>1</th>
<th>3</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human only</td>
<td>55.91</td>
<td>54.05</td>
<td>47.81</td>
<td>40.14</td>
</tr>
<tr>
<td>LR-ID</td>
<td>54.95</td>
<td>49.77</td>
<td>37.71</td>
<td><strong>25.36</strong></td>
</tr>
<tr>
<td>GD-ID</td>
<td>57.51</td>
<td>53.52</td>
<td>44.05</td>
<td>36.99</td>
</tr>
<tr>
<td>Corp-ID</td>
<td>54.19</td>
<td>49.37</td>
<td>39.06</td>
<td>29.54</td>
</tr>
<tr>
<td>Web-ID</td>
<td><strong>51.88</strong></td>
<td><strong>48.32</strong></td>
<td><strong>37.49</strong></td>
<td><strong>29.49</strong></td>
</tr>
</tbody>
</table>

ID: Interactive Decoding
WER of corrections for initial transcriptions of WER above 40%.

<table>
<thead>
<tr>
<th># c/segment</th>
<th>1</th>
<th>3</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human only</td>
<td>55.91</td>
<td>54.05</td>
<td>47.81</td>
<td>40.14</td>
</tr>
<tr>
<td>LR-ID</td>
<td>54.95</td>
<td>49.77</td>
<td>37.71</td>
<td>25.36</td>
</tr>
<tr>
<td>GD-ID</td>
<td>57.51</td>
<td>53.52</td>
<td>44.05</td>
<td>36.99</td>
</tr>
<tr>
<td>Corp-ID</td>
<td>54.19</td>
<td>49.37</td>
<td>39.06</td>
<td>29.54</td>
</tr>
<tr>
<td>Web-ID</td>
<td>51.88</td>
<td>48.32</td>
<td>37.49</td>
<td>29.49</td>
</tr>
</tbody>
</table>

ID: Interactive Decoding
Interactive decoding conclusion

- Effectiveness of interactive strategies
- Global cost reducing
- Driving methods:
  - Graph density is rather inefficient
  - Left-Right is the best way to produce a perfect transcription
  - Semantic methods are effective for massively erroneous transcriptions
- Improvement of the semantic quality using semantic strategies
- Efficient way of correcting transcriptions dedicated to:
  - speech indexing
  - speech understanding
Thanks you for your attention!