Automatic Emphatic Information Extraction from Aligned Acoustic Data and Its Application on Sentence Compression

Yanju Chen and Rong Pan
School of Data and Computer Science
Sun Yat-sen University
Intro.: Tell A Story for Kids

• From *Sleeping Beauty*:
  • *Oh, how happy they were!*

• *They shared their joy by inviting seven wise fairies to the palace.*

• *Now there was one other fairy whose magic was more powerful than all the wise ones put together.*
Intro.: Tell A Story for Kids

• From Sleeping Beauty:
  • Oh, how happy they were!

• They shared their joy by inviting seven wise fairies to the palace.

• Now there was one other fairy whose magic was more powerful than all the wise ones put together.
Motivations

• To Extract the Semantic/Prosody Information
  • From Some Acoustic Patterns in Speech
  • From Eye Tracking (Klerke etc. 2016)
• To Represent/Model Semantic Patterns in Speech
• To Utilize Extracted Semantic Information
We use speeches to generate labels for texts, and use texts to predict these labels, and incorporate the patterns in texts into sentence compression task.
Intro.: Prosody in Speech

- Emphasized Semantic Information
  - Uncertainty
  - Contrast
  - etc

- Perceivable by Listeners (Prosody Detection)
  - Lower-Level Acoustic Features
  - Higher-Level Acoustic Features
Intro.: Prosody *Prediction*

- Predict Prosodic Prominence from *Lexical Features Only*
  - Word-Based Prosodic Patterns
  - Manual Text-To-Speech Alignment
  - Hand-Crafted Lexical/Semantic Features

- Related Works
  - (Brenier etc. 2005): Maxent, Read Text
  - (Brenier 2008): More Advanced Lexical/Semantic Features, Read & Speech Texts
Intro.: Prosody Application

• Text-To-Speech Synthesis

• Related Applications:
  • Emotion Detection (Cao etc. 2014)
  • Disfluency Detection (Ferguson etc. 2015)
  • Deception Detection (Levitan etc. 2016)
  • Speaker State Detection (Wang etc. 2013)
Challenges in Prosody Prediction

- Large-Scale Annotation
  - Large-Scale Speech Data with Transcriptions
- High Labeling Cost: Sometimes Unaffordable
- Normalization
Our Solution: Prosody Prediction

• Weak Supervision
  • **Automatic Speech-To-Text Alignment**
    -> Large-Scale Data
  • **Empirical Rules**
    -> Weakly/Noisily Labeled Data
• **Using Distinctive Acoustic Features**
  -> Normalization
Intro.: Sentence Compression

• Target: Generate Shorter Paraphrases

• Application
  • Automatic Summarization
  • Assistive Applications

• Extractive (Deletion-Based) Sentence Compression
  • Generate Subsequences of the Input Sequences
Challenges in Sentence Compression

• Relying Heavily on Manual Syntactic Information
  • Vulnerability in Error Propagation
  • Manual Labeling Required Training Syntactic Parsers

• Incorporation of Extra Data/Supervision
  • How to generate large-scale extra data
Our Solution: Sentence Compression

• Multitask & Extra Data
  • **Lexical Prosody Dataset**
    -> Get rid of manual labeling
  • **Multitask Learning**
    -> Incorporate prosody in learning
The Problems

- Where to find large-scale aligned acoustic data
- How to generate prosodic representation for every word automatically
- How to utilize the labeled data and incorporate them into Sentence Compression task
Our Solution

Prosody Dataset Construction

• Source: Lit2Go (Audio Book)
• Aligner: cmusphinx
• Feature: Standard Word Duration

\[ S.Duration = \frac{\text{total time duration}}{\text{num. of syllables}} \]

• Normalization (within a sentence)
• Categorization
Our Solution

Prosody Dataset Construction: Algorithm & Example

Word | Jim | was | laid | up | four | days | and | nights.
--- | --- | --- | --- | --- | --- | --- | --- | ---
Duration(ms) | 206 | 278 | 348 | 148 | 417 | 274 | 260 | 500
Emp.Lvl. (30) | 18 | 12 | 20 | 30 | 22 | 26 | 11 | 21
Emp.Lvl. | E | E | E | E | E | E | E | E
Emp.Lvl. (2) | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1
Our Solution
Prosody Dataset Construction: Example

• Jim was *laid up* for *four days* and *nights*.
  • (EP: 001101101)

• But it was *too dark* to *see* yet, *so* we *made* the *canoe* *fast* and *set in* her *to* wait for *daylight*.
  • (EP: 00011010,10101101101001)

• I didn’t need anybody to *tell me* that *that* was an *awful bad sign* and would *fetch me* some *bad luck*, so I was *scared* and most *shook* the *clothes off* of *me*.
  • (EP: 000001101001110011011,000100101101)
Prosody Dataset Details

- Basic Information of Collected Acoustic Data
- Every sentence is a sample in the dataset.

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>authors:</td>
<td>208</td>
<td>books:</td>
<td>205</td>
</tr>
<tr>
<td>genres:</td>
<td>22</td>
<td>passages:</td>
<td>4198</td>
</tr>
<tr>
<td>sentences:</td>
<td>286,083</td>
<td>words:</td>
<td>5,881,720</td>
</tr>
<tr>
<td>vocab. size:</td>
<td>48,204</td>
<td>mean sent. len.:</td>
<td>20</td>
</tr>
</tbody>
</table>
Our Solution
Modeling Prosodic Patterns: Settings

- Problem Type: Sequence Labeling
- Architecture: LSTM, Bi-LSTM
  \[ \theta^* = \arg \max_{\theta} \sum_{X,A} \log p(A|X; \theta), \quad \hat{A} = \arg \max_{A} p(A|X; \theta^*) \]
- Evaluation Metrics:
  - Word-Based Accuracy, Sentence-Based Accuracy
- Dataset Characteristics:
  - 230k(train), 25k(valid), 28k(test)
- Results:
  - LSTM-(82.90%, 9.47%), Bi-LSTM-(85.24%, 14.42%)
Our Solution

Multi-Task Sentence Compression

• Problem Type: Multi-Task Sequence Labeling
• Architectures: LSTM, Bi-LSTM, Stacked Bi-LSTM

Figure 1: Basic LSTM Unrolled Through Time
Our Solution
Multi-Task Sentence Compression

- Training: Alternative Multi-Task
- Compression Datasets: GOOGLE, BROADCAST
- Extra Datasets: Pre-Processed Eye-Tracking Dataset
- Evaluation Metrics:
  - W.Acc: Word-Based Accuracy
  - S.Acc: Sentence-Based Accuracy
  - $F1_0$: F1-Scores of Label 0
  - $F1_1$: F1-Scores of Label 1
- Without Any Extra Syntactic/Semantic Information
## Results & Analysis: GOOGLE

### Table 5: Performance on GOOGLE Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>GOOGLE</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>W.Acc</td>
<td>F1₀</td>
<td>F1₁</td>
<td>S.Acc</td>
</tr>
<tr>
<td>LSTM</td>
<td>Baseline</td>
<td>79.15</td>
<td>82.73</td>
<td>73.70</td>
<td>6.51</td>
</tr>
<tr>
<td></td>
<td>Emphatic</td>
<td><strong>79.40</strong></td>
<td><strong>82.86</strong></td>
<td><strong>74.19</strong></td>
<td>7.18</td>
</tr>
<tr>
<td></td>
<td>Gaze.fp</td>
<td>79.27</td>
<td>82.84</td>
<td>73.81</td>
<td>6.58</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>Baseline</td>
<td>79.79</td>
<td>83.31</td>
<td>74.40</td>
<td>7.19</td>
</tr>
<tr>
<td></td>
<td>Emphatic</td>
<td><strong>80.14</strong></td>
<td><strong>83.73</strong></td>
<td><strong>74.50</strong></td>
<td><strong>8.11</strong></td>
</tr>
<tr>
<td></td>
<td>Gaze.fp</td>
<td>79.71</td>
<td>83.40</td>
<td>73.97</td>
<td>7.53</td>
</tr>
<tr>
<td>Stacked Bi-LSTM</td>
<td>Baseline</td>
<td>79.94</td>
<td>83.40</td>
<td>74.63</td>
<td>8.00</td>
</tr>
<tr>
<td></td>
<td>Emphatic</td>
<td><strong>80.30</strong></td>
<td><strong>83.74</strong></td>
<td><strong>74.99</strong></td>
<td><strong>9.26</strong></td>
</tr>
<tr>
<td></td>
<td>Gaze.fp</td>
<td>79.95</td>
<td>83.48</td>
<td>74.50</td>
<td>8.53</td>
</tr>
</tbody>
</table>
## Results & Analysis: BROADCAST1

Table 6: Performance on BROADCAST1 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>(W.\text{Acc})</th>
<th>(F_{10})</th>
<th>(F_{11})</th>
<th>(S.\text{Acc})</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>Baseline</td>
<td>72.28</td>
<td>14.56</td>
<td>83.43</td>
<td>10.93</td>
</tr>
<tr>
<td></td>
<td><strong>Emphatic</strong></td>
<td><strong>72.70</strong></td>
<td><strong>19.37</strong></td>
<td><strong>83.53</strong></td>
<td>10.87</td>
</tr>
<tr>
<td></td>
<td>Gaze.fp</td>
<td>72.69</td>
<td>18.56</td>
<td><strong>83.56</strong></td>
<td>10.90</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>Baseline</td>
<td>72.76</td>
<td>21.17</td>
<td>83.51</td>
<td>11.30</td>
</tr>
<tr>
<td></td>
<td><strong>Emphatic</strong></td>
<td><strong>73.56</strong></td>
<td><strong>25.93</strong></td>
<td><strong>83.87</strong></td>
<td><strong>11.95</strong></td>
</tr>
<tr>
<td></td>
<td>Gaze.fp</td>
<td>73.34</td>
<td>23.98</td>
<td>83.82</td>
<td>11.81</td>
</tr>
</tbody>
</table>
Our Solution

Results & Analysis: BROADCAST2

Table 7: Performance on BROADCAST2 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>W.Acc</th>
<th>F1₀</th>
<th>F1₁</th>
<th>S.Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>Baseline</td>
<td>79.10</td>
<td>13.27</td>
<td>88.12</td>
<td>22.19</td>
</tr>
<tr>
<td>LSTM</td>
<td>Emphatic</td>
<td>79.34</td>
<td>17.20</td>
<td>88.19</td>
<td>22.25</td>
</tr>
<tr>
<td>LSTM</td>
<td>Gaze.fp</td>
<td>79.42</td>
<td>15.98</td>
<td>88.27</td>
<td>22.12</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>Baseline</td>
<td>79.78</td>
<td>22.89</td>
<td>88.35</td>
<td>22.82</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>Emphatic</td>
<td>80.37</td>
<td>26.60</td>
<td>88.66</td>
<td>23.19</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>Gaze.fp</td>
<td>80.24</td>
<td>26.11</td>
<td>88.59</td>
<td>22.97</td>
</tr>
</tbody>
</table>
Our Solution

Results & Analysis:
BROADCAST3

Table 8: Performance on BROADCAST3 Dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>Data</th>
<th>W.Acc</th>
<th>$F_{10}$</th>
<th>$F_{11}$</th>
<th>S.Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>Baseline</td>
<td>66.85</td>
<td>36.22</td>
<td>77.55</td>
<td>9.60</td>
</tr>
<tr>
<td></td>
<td>Emphatic</td>
<td>67.06</td>
<td>37.93</td>
<td>77.52</td>
<td>9.70</td>
</tr>
<tr>
<td></td>
<td>Gaze.fp</td>
<td>67.19</td>
<td><strong>40.38</strong></td>
<td>77.34</td>
<td>8.56</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>Baseline</td>
<td>67.58</td>
<td>38.94</td>
<td>77.86</td>
<td>11.48</td>
</tr>
<tr>
<td></td>
<td>Emphatic</td>
<td>68.35</td>
<td>38.39</td>
<td><strong>78.66</strong></td>
<td><strong>11.65</strong></td>
</tr>
<tr>
<td></td>
<td>Gaze.fp</td>
<td>68.23</td>
<td>38.01</td>
<td>78.59</td>
<td>11.57</td>
</tr>
</tbody>
</table>
Future Works

- Better Tuned Extraction
  - Speaker Normalization
  - More Acoustic Features (f0 & Intensity)
- More Sophisticated Multitask Training
- Incorporation with More NLP Tasks
  - Low-Level: POS Tagging, NER, SRL, ...
  - High-Level: Sentiment, QA, Translation, ...
Thank you!