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.*

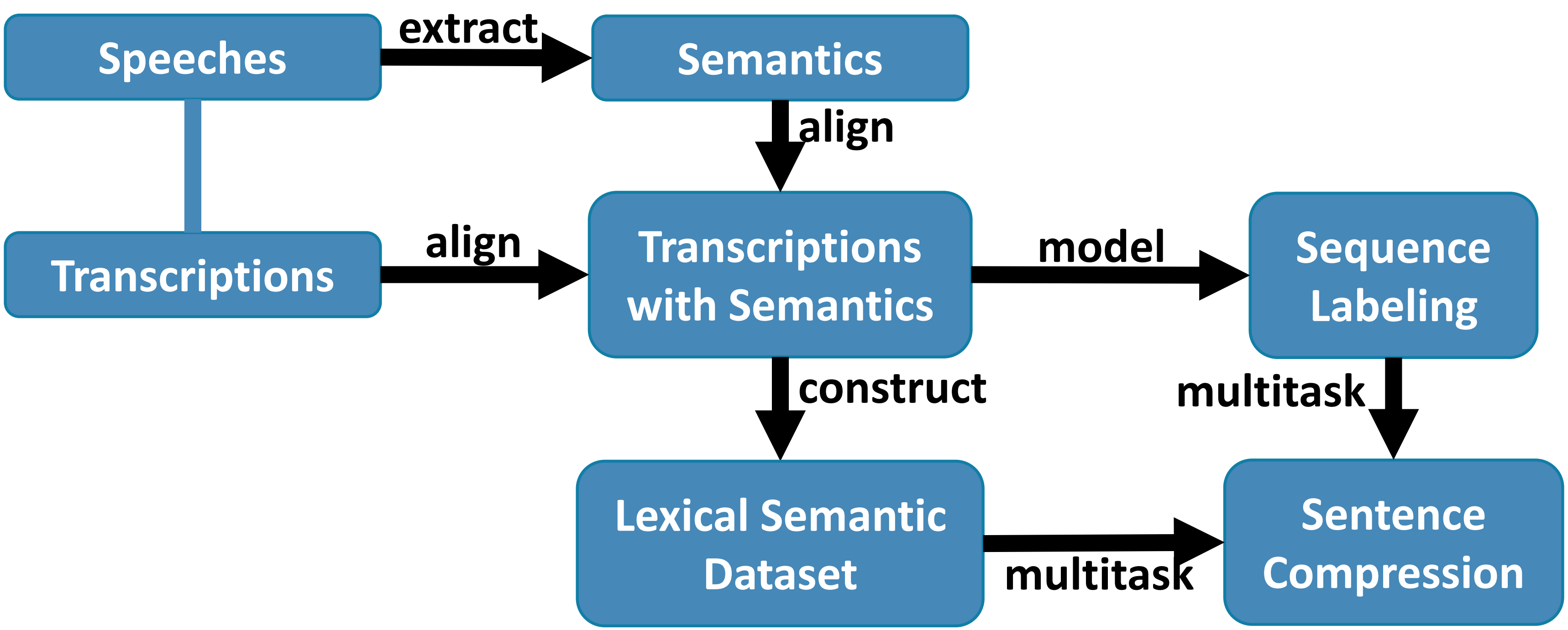
Intro.: Tell A Story for Kids

- From *Sleeping Beauty*:
  - *Oh, how **happy** they were!*
  - *They shared their joy <break> by inviting **seven wise fairies** <break> to the palace.*
  - *Now <break> there was one other **fairy** <break> 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

Our Work



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

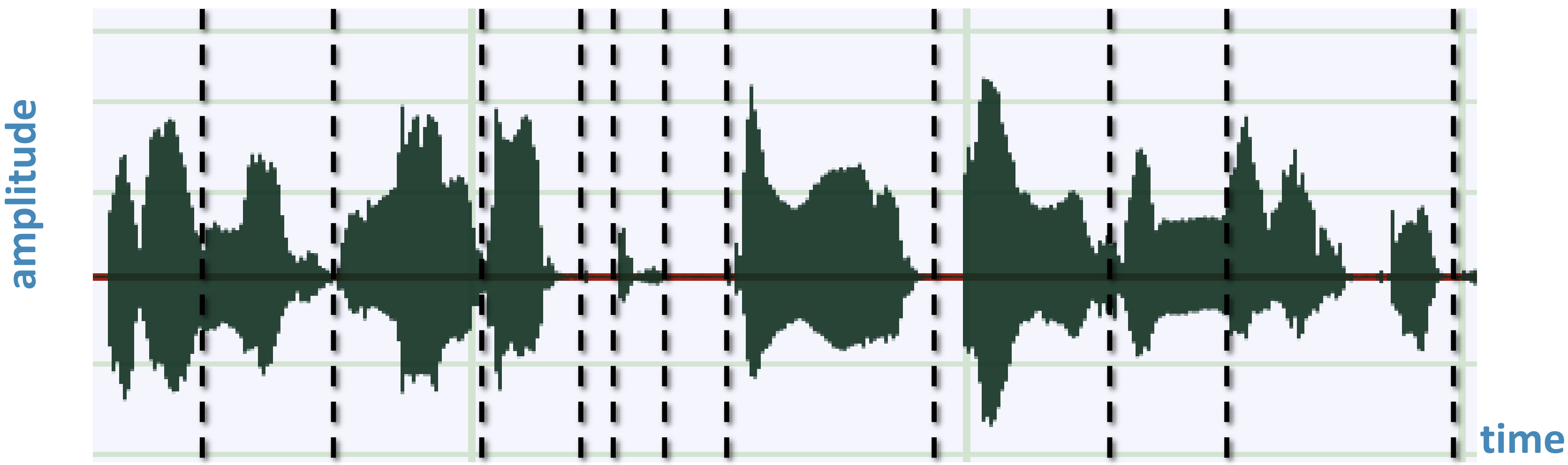
- Source: *Lit2Go* (Audio Book)
- Aligner: cmusphinx
- Feature: Standard Word Duration

$$S.Duration = \frac{total\ time\ duration}{num.\ of\ syllables}$$

- 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
Emp.Lvl. (2)	0	0	1	1	0	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)

Our Solution

Prosody Dataset Details

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

Table 1: Basic Information of Collected Acoustic Data

authors:	208	books:	205
genres:	22	passages:	4198
sentences:	286,083	words:	5,881,720
vocab. size:	48,204	mean sent. len.:	20

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

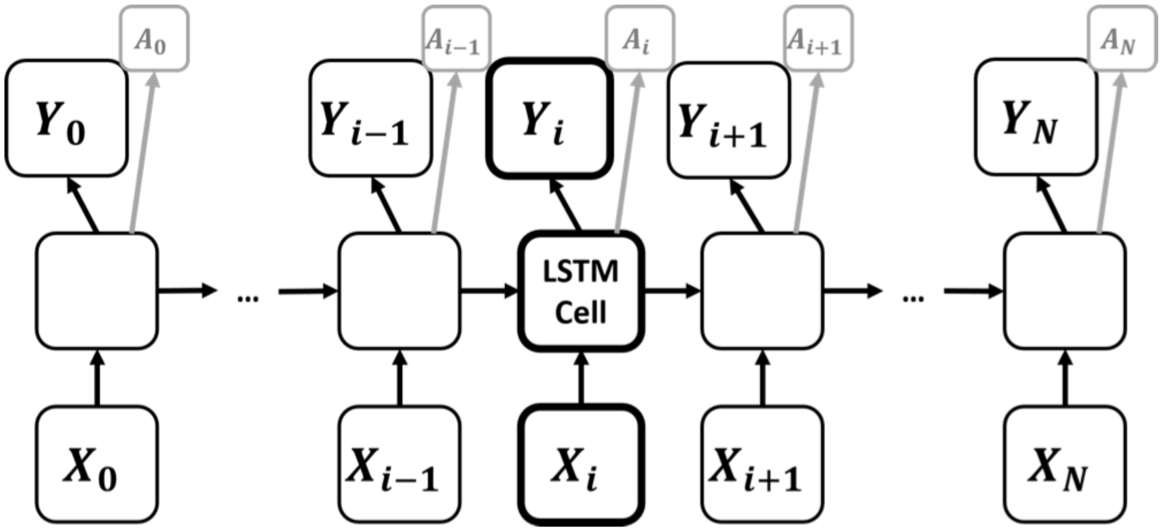


Figure 1: Basic LSTM Unrolled Through Time

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**

Our Solution

Results & Analysis: GOOGLE

Table 5: Performance on GOOGLE Dataset

Model	Data	GOOGLE			
		$W.Acc$	$F1_0$	$F1_1$	$S.Acc$
LSTM	Baseline	79.15	82.73	73.70	6.51
	Emphatic	**79.40**	**82.86**	**74.19**	**7.18**
	Gaze.fp	79.27	82.84	73.81	6.58
Bi-LSTM	Baseline	79.79	83.31	74.40	7.19
	Emphatic	**80.14**	**83.73**	**74.50**	**8.11**
	Gaze.fp	79.71	83.40	73.97	7.53
Stacked Bi-LSTM	Baseline	79.94	83.40	74.63	8.00
	Emphatic	**80.30**	**83.74**	**74.99**	**9.26**
	Gaze.fp	79.95	83.48	74.50	8.53

Our Solution

Results & Analysis: BROADCAST1

Table 6: Performance on BROADCAST1 Dataset

Model	Data	BROADCAST1			
		$W.Acc$	$F1_0$	$F1_1$	$S.Acc$
LSTM	Baseline	72.28	14.56	83.43	**10.93**
	Emphatic	**72.70**	**19.37**	83.53	10.87
	Gaze.fp	72.69	18.56	**83.56**	10.90
Bi-LSTM	Baseline	72.76	21.17	83.51	11.30
	Emphatic	**73.56**	**25.93**	**83.87**	**11.95**
	Gaze.fp	73.34	23.98	83.82	11.81

Our Solution

Results & Analysis: BROADCAST2

Table 7: Performance on BROADCAST2 Dataset

Model	Data	BROADCAST2			
		$W.Acc$	$F1_0$	$F1_1$	$S.Acc$
LSTM	Baseline	79.10	13.27	88.12	22.19
	Emphatic	79.34	**17.20**	88.19	**22.25**
	Gaze.fp	**79.42**	15.98	**88.27**	22.12
Bi-LSTM	Baseline	79.78	22.89	88.35	22.82
	Emphatic	**80.37**	**26.60**	**88.66**	**23.19**
	Gaze.fp	80.24	26.11	88.59	22.97

Our Solution

Results & Analysis: BROADCAST3

Table 8: Performance on BROADCAST3 Dataset

Model	Data	BROADCAST3			
		$W.Acc$	$F1_0$	$F1_1$	$S.Acc$
LSTM	Baseline	66.85	36.22	**77.55**	9.60
	Emphatic	67.06	37.93	77.52	**9.70**
	Gaze.fp	**67.19**	**40.38**	77.34	8.56
Bi-LSTM	Baseline	67.58	**38.94**	77.86	11.48
	Emphatic	**68.35**	38.39	**78.66**	**11.65**
	Gaze.fp	68.23	38.01	78.59	11.57

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!