

A review of NLP research work of Taylor Berg-Kirkpatrick

Prepared by: Ritesh Sarkhel



Biography

B.S. : University of California, Berkley


PhD : University of California, Berkley

Intern: Machine Translation, Google

Faculty: CMU, since 2016

Research Interests

- Natural language processing and machine learning, using unsupervised methods for deciphering hidden structure.
- End applications include: various types of human artifacts, including natural language and diverse sources like early modern books, handwritten text, historical ciphers, and music.



Learning Bilingual Lexicons from Monolingual Corpora

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick and Dan Klein

ACL '08

Motivation

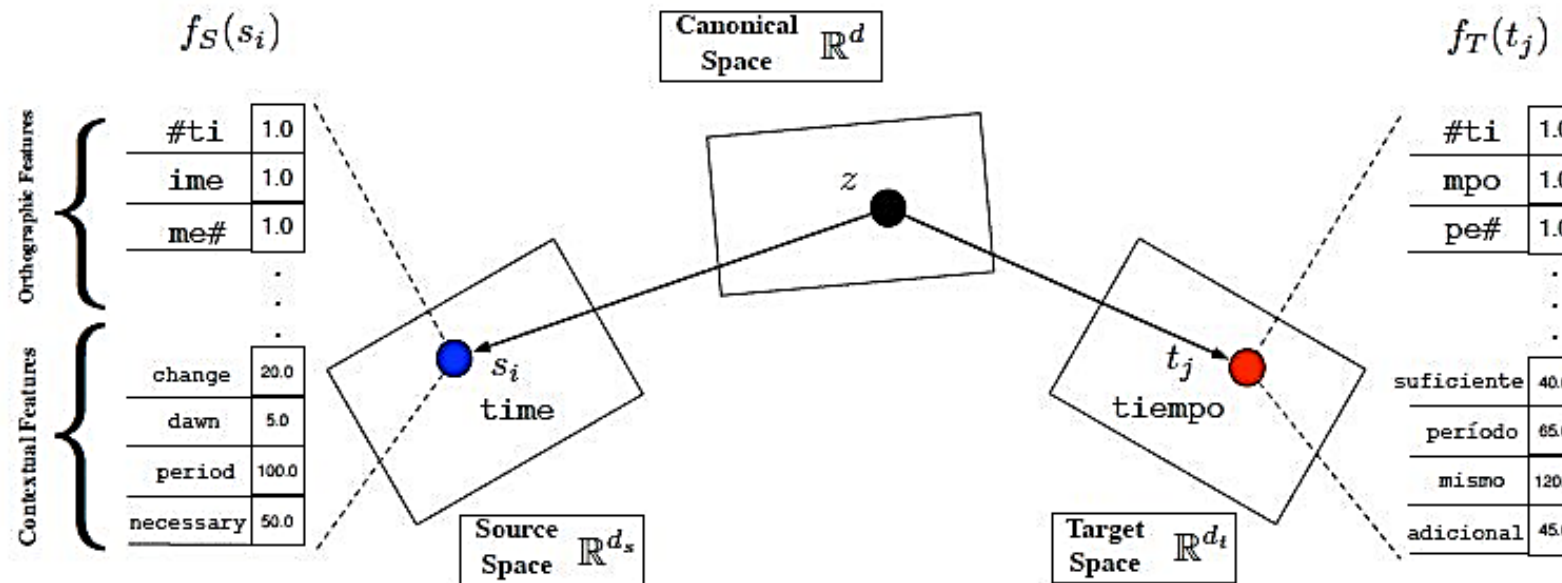
- Although parallel text is plentiful for some language pairs such as English-Chinese or English-Arabic, it is scarce or even non-existent for most others, such as English-Hindi or French-Japanese
- Parallel text could be scarce for a language pair even if monolingual data is readily available for both languages.
- Objective: *Generate translation pairs from monolingual corpus using a generative model.*

Methodology

- $S = \{s_1, s_2, \dots, s_n\}$: Source corpus of n source words
- $T = \{t_1, t_2, \dots, t_m\}$: Target corpus of m target words
- Output: $m = \{(s_i, t_j), \forall i, j\}$
- In other words: Find optimal full bipartite matching between S and T.

Methodology (contd.)

- Initialize the matching prior as uniform distribution
- For each matched pair $\{s_i, t_j\}$ extract feature set $f_S(s_i)$ and $f_T(t_j)$
- ‘Explain away’ translation pairs in a language independent canonical subspace



Methodology (contd.)

- $f_s(s_i) \sim \text{Multivariate Gaussian}(W_s z_{ij}, \psi_s)$
- $f_t(t_j) \sim \text{Multivariate Gaussian}(W_t z_{ij}, \psi_t)$

- Maximize the likelihood of :

$$l(\theta) = \log \sum_m p(m, s, t; \theta)$$

- $\theta = \{W_s, W_t, \psi_s, \psi_t\}$
- Approximate $p(m, s, t; \theta) = \sum_{(i,j)} w_{ij} + C$
- Optimize θ using a modified EM algorithm.

Experimental Results

Setting	$p_{0.1}$	$p_{0.25}$	$p_{0.33}$	$p_{0.50}$	Best- F_1
EDITDIST	58.6	62.6	61.1	—	47.4
ORTHO	76.0	81.3	80.1	52.3	55.0
CONTEXT	91.1	81.3	80.2	65.3	58.0
MCCA	87.2	89.7	89.0	89.7	72.0

Table 1: Performance of EDITDIST and our model with various features sets on EN-ES-W. See section 5.

Setting	$p_{0.1}$	$p_{0.25}$	$p_{0.33}$	$p_{0.50}$	Best- F_1
EN-ES-G	75.0	71.2	68.3	—	49.0
EN-ES-W	87.2	89.7	89.0	89.7	72.0
EN-ES-D	91.4	94.3	92.3	89.7	63.7
EN-ES-P	97.3	94.8	93.8	92.9	77.0



Unsupervised Transcription of Piano Music

Taylor Berg-Kirkpatrick Jacob Andreas Dan Klein

NIPS '14

Motivation

- Probabilistic model that describes the process by which discrete musical events give rise to (separate) acoustic signals for each keyboard note, and the process by which these signals are superimposed to produce the observed data.
- Output: Given a piano recording, without any previously seen data, the model generates a MIDI like symbolic representation of the audio.

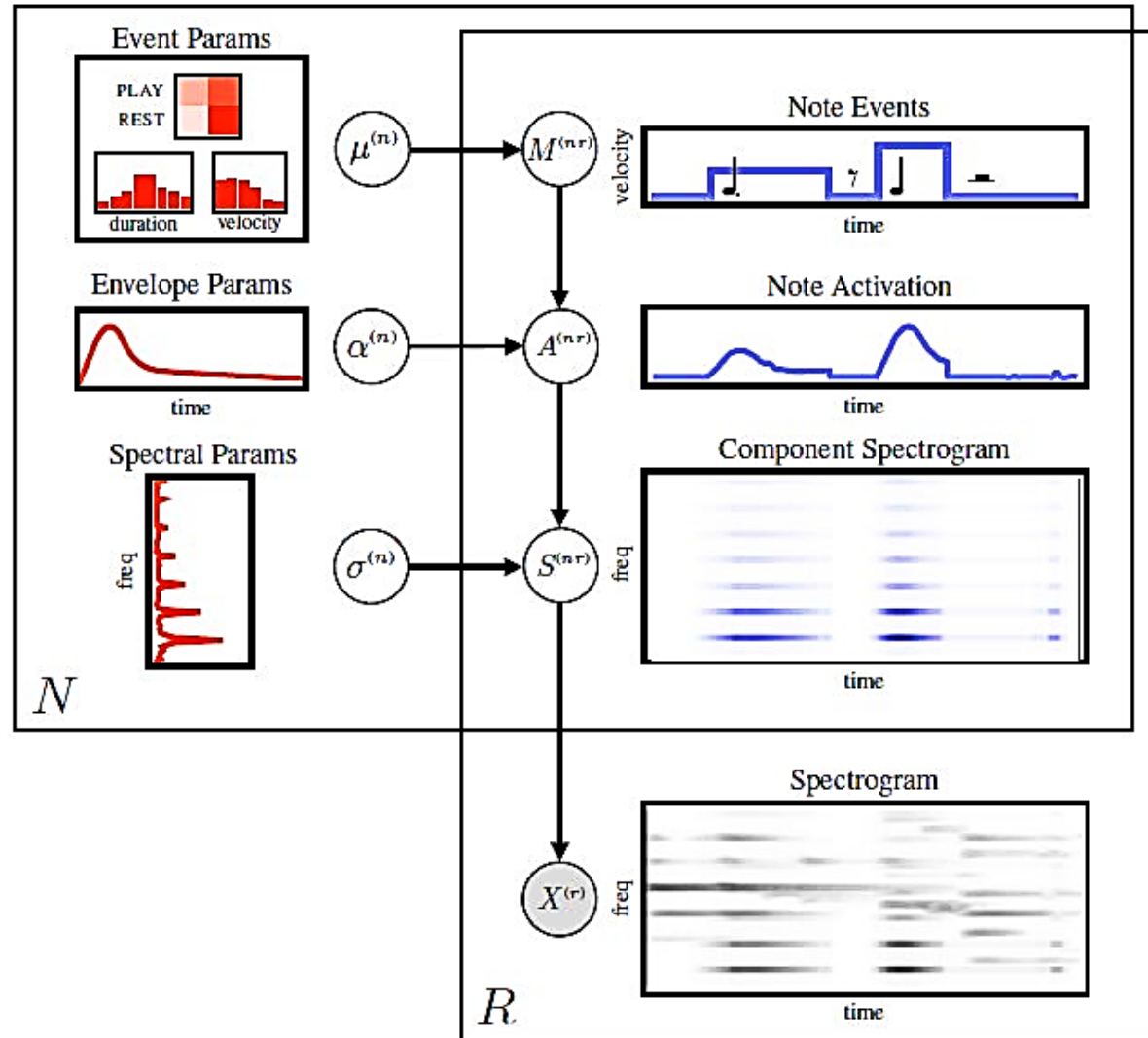
Why is this task difficult?

- Even individual piano notes are quite rich.
 - A single note is not simply a fixed-duration sine wave at an appropriate frequency, a full spectrum of harmonics that rises and falls in intensity.
 - Profiles vary from piano to piano and therefore must be learned in a recording-specific way => supervised way.
- Piano music is generally polyphonic, i.e. multiple notes are played simultaneously.
 - Combinations of notes exhibit ambiguous harmonic collisions
 - Inherent source separation problem.

Why is this task difficult? (contd.)

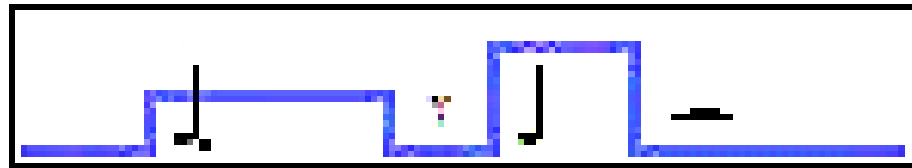
- Most previous work:
 - Better modelling of the discrete musical structure
 - Or, better adapting to the timbral properties of the source instrument
 - Why?
 - Coupling these discrete models with timbral adaptation and source separation breaks the conditional independence assumptions that the dynamic programs (e.g. HMM, Semi-markov models) rely on.
- Tackles these discrete and timbral modelling problems jointly
 - New generative model that reflects the causal process underlying piano sound generation
 - Tractable approximation to the inference problem over transcriptions and timbral parameters

Model



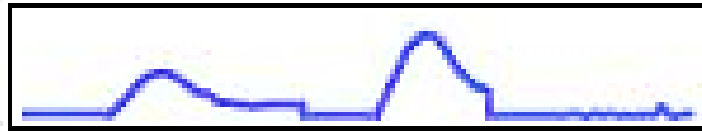
Model (contd.)

- Consider a song S , divided into T time steps. The transcription will be l musical events long.
- The component model for a single note C' in S has 3 primary random variables:
 - M , a sequence of l symbolic musical events, analogous to the locations and values of symbols along the C' in sheet music,

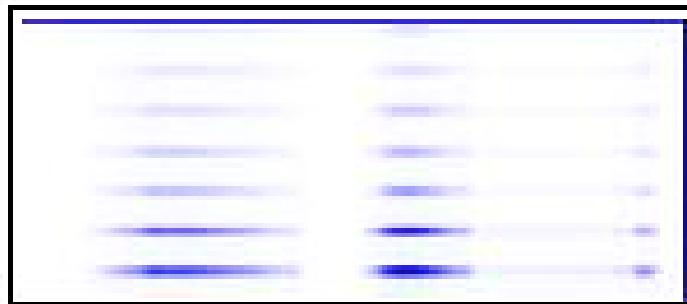


Model(contd.)

- A, a time series of T activations, analogous to the loudness of sound emitted by the C' piano string over time as it peaks and attenuates during each event in M.



- S, a spectrogram of T frames, specifying the spectrum of frequencies over time in the acoustic signal produced by the C' string.



Model(contd.)

- Joint distribution of a note is:

$$P(S, A, M | \sigma^{C'}, \alpha^{C'}, \mu^{C'}) = P(M | \mu^{C'}) * P(A | M, \alpha^{C'}) * P(S | A, \sigma^{C'})$$

- $\mu^{C'}$ = How long the C' string is likely to be held for (duration), and how hard it is likely to be pressed (velocity).
- $\alpha^{C'}$ = The shape of the rise and fall of the string's activation each time the note is played.
- $\sigma^{C'}$ = The frequency distribution of sounds produced by the C' string

Full model of a song

- Each pair of note n (on a standard piano 88 notes) and song r , is defined by:
 - Musical events model ($\mathbf{M}^{nr} = \{m^{1r}, m^{2r}, \dots m^{nr}\}$)
 - Activation model ($\mathbf{A}^{nr} = \{a^{1r}, a^{2r} \dots a^{nr}\}$)
 - Spectrogram model ($\mathbf{S}^{nr} = \{s^{1r}, s^{2r} \dots s^{nr}\}$)
 - Event parameters ($\boldsymbol{\mu}^n = \{\mu^1, \mu^2 \dots \mu^n\}$)
 - Activation parameters ($\boldsymbol{\alpha}^n = \{\alpha^1, \alpha^2 \dots \alpha^n\}$)
 - Spectrogram parameters ($\boldsymbol{\sigma}^n = \{\sigma^1, \sigma^2 \dots \sigma^n\}$)

Learning and Inference

- Goal: Estimate the unobserved musical events for each song, $M(r)$, as well as the unknown envelope and spectral parameters of the piano that generated the data, σ and α .
 - Compute the posterior distribution on M , σ and α .

$$\max_{\bar{M}, \bar{A}, \alpha, \sigma} \prod_r \left[\sum_{S(r)} P(X^{(r)}, S^{(r)}, A^{(r)}, M^{(r)} | \mu, \alpha, \sigma) \right] \cdot P(\alpha, \sigma)$$

- Approximate the joint MAP estimates of M , A , σ and α via iterated conditional modes by marginalizing over the component spectrograms S .
- Update parameters via block-coordinate ascent.

Experimental Results

- Evaluated on MIDI-Aligned Piano Sounds (MAPS) corpus.
 - First 30 seconds of each of the 30 ENSTDkAm recordings as a development set
 - First 30 seconds of each of the 30 ENSTDkCl recordings as a test set.
- Symbolic music data from the IMSLP library used to estimate the event parameters in the model.

Experimental Results(contd.)

- State of the art results
- > 10% improvement over best published result

System	Onsets			Frames		
	P	R	F ₁	P	R	F ₁
Discriminative [7]	76.8	65.1	70.4	-	-	-
Benetos [2]	-	-	68.6	-	-	68.0
Vincent [3]	62.7	76.8	69.0	79.6	63.6	70.7
O’Hanlon [4]	48.6	73.0	58.3	73.4	72.8	73.2
This work	78.1	74.7	76.4	69.1	80.7	74.4

Table 1: Unsupervised transcription results on the MAPS corpus. “Onsets” columns show scores for identification (within ± 50 ms) of note start times. “Frames” columns show scores for 10ms frame-level evaluation. Our system achieves state-of-the-art results on both metrics.^[2]

Questions?



Firm