A Statistical Approach to Understanding Natural Language

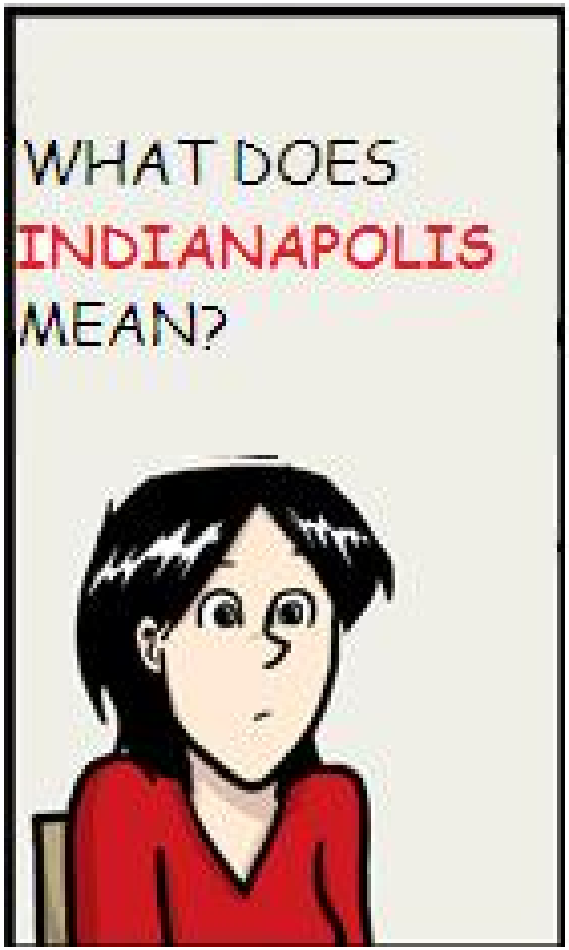
Dongyu Lin[1]
(Joint work with Dean P. Foster[1], Sham M. Kakade[2], Lyle H. Ungar[3])

[1]Department of Statistics, The Wharton School, University of Pennsylvania
[2]Toyota Technological Institute at Chicago
[3]Department of Computer and Information Science, University of Pennsylvania

Statistics Workshop at Yale University
Innovation and Inventiveness in Statistics Methodologies
May 14-17, 2009

Word Sense Disambiguation

Dongyu was reading an article...



Saturday, Nov. 8, 2008

Ashes of USS Indianapolis survivor buried at sea

By ERIC TALMADGE
The Associated Press

YOKOSUKA, Kanagawa Pref. — When the submarine USS Ohio surfaced at sea and Machinist Mate 1st Class Jason Witty emerged from the hatch to look around, he saw calm, blue water under a peaceful sky — perfect for the solemn task he was about to perform.

On the map, the Ohio was afloat in just another indistinguishable expanse of the Pacific Ocean. As Witty stood on deck holding a silver pitcher, the vessel was alone.

Just like the ill-fated USS Indianapolis, 63 years earlier.

Word Sense Disambiguation

Dongyu is nonnative, so she looked up WikipediA...

Indianapolis (disambiguation)

From Wikipedia, the free encyclopedia

Indianapolis is the capital and largest city of the U.S. state of Indiana.

Indianapolis may also refer to:

Places in the United States

- Indianapolis (balance), Indiana, a U.S. Census Bureau designation corresponding to that portion of the consolidated city-county entity that does not include any of the other incorporated places within the boundary
- Indianapolis, Iowa, in Mahaska County, roughly between Iowa City and Des Moines

Other

- USS *Indianapolis*, the name of three United States Navy ships
- Indianapolis Motor Speedway near Indianapolis, Indiana
- Indianapolis 500, famous annual motor race at the speedway
- Indianapolis Tennis Championships, an ATP event that is part of the U.S. Open Series

Word Sense Disambiguation

Dongyu wished her machine was smarter enough to disambiguate...

The pitcher contained the ashes of Witty's grandfather, Boatswain Mate 2nd Class Eugene Morgan, who had survived the sinking of the Indianapolis — one of the worst tragedies for the U.S. Navy in World War II.

Morgan had died of a heart attack in June at age 87, just before Witty went to sea, and among his last wishes was the desire to be rejoined with his shipmates at roughly the same spot in the Pacific where the Indianapolis went down.

Witty, sitting in a wardroom of the Ohio at the U.S. Navy base in Yokosuka, recounted the Oct. 2 burial at sea, saying he had never participated in one before.

He had sheepishly asked one of the officers if his grandfather's wish could be

Understanding Natural Language

Our Long Term Goal

To develop a software that

- in a specific sense "understands" natural language
- is able to offer the definition of a word within the context

Understanding Natural Language

Our Long Term Goal

To develop a software that

- in a specific sense "understands" natural language
- is able to offer the definition of a word within the context

Three Steps

1. Disambiguation
2. Translation
3. Reference

Understanding Natural Language

Our Long Term Goal

To develop a software that

- in a specific sense "understands" natural language
- is able to offer the definition of a word within the context

Three Steps

1. Disambiguation

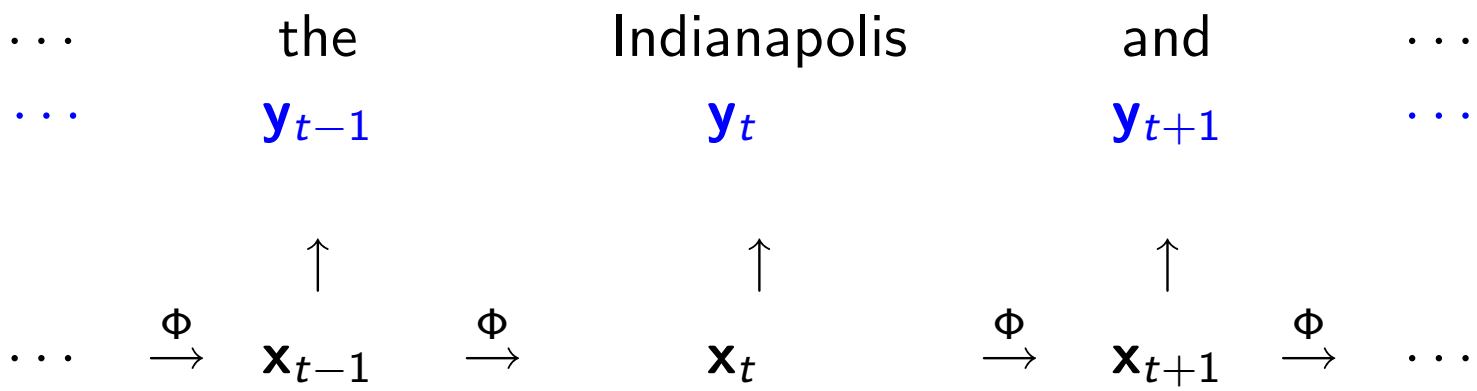
The State-Space Model

- $\mathcal{D} = \{w_1, \ldots, w_d\}$ is a finite dictionary
- Document $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$, with $\mathbf{y}_t \in \{0,1\}^d$, $y_t^j = I_{\{w_j = v_t\}}$, where v_t is the t^{th} word in the document.
- $\mathbf{x}_t \in \mathbb{R}^m$ hidden states
- we have the following state-space model

$$\begin{aligned} \mathbf{y}_t &= A\mathbf{x}_t + \mathbf{u}_t, \\ \mathbf{x}_{t+1} &= \Phi\mathbf{x}_t + \boldsymbol{\delta}_t. \end{aligned} \tag{1}$$

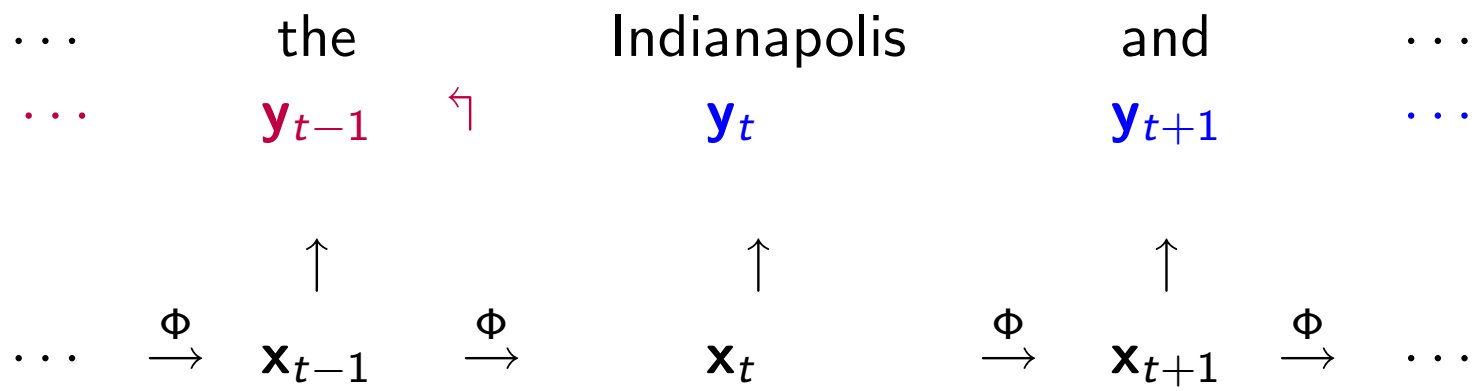
- $A \in \mathbb{R}^{d\times m}$, $\Phi \in \mathbb{R}^{m\times m}$
- $\mathbf{u}_t \in \mathbb{R}^d$, $\boldsymbol{\delta}_t \in \mathbb{R}^m$ white noise sequences, with $\mathrm{Var}(\mathbf{u}_t) = Q$, $\mathrm{Var}(\mathbf{v}_t) = R$, $\mathbb{E}\mathbf{u}_t\mathbf{v}_s' = S\delta_{ts}$.

A CCA approach



A CCA approach

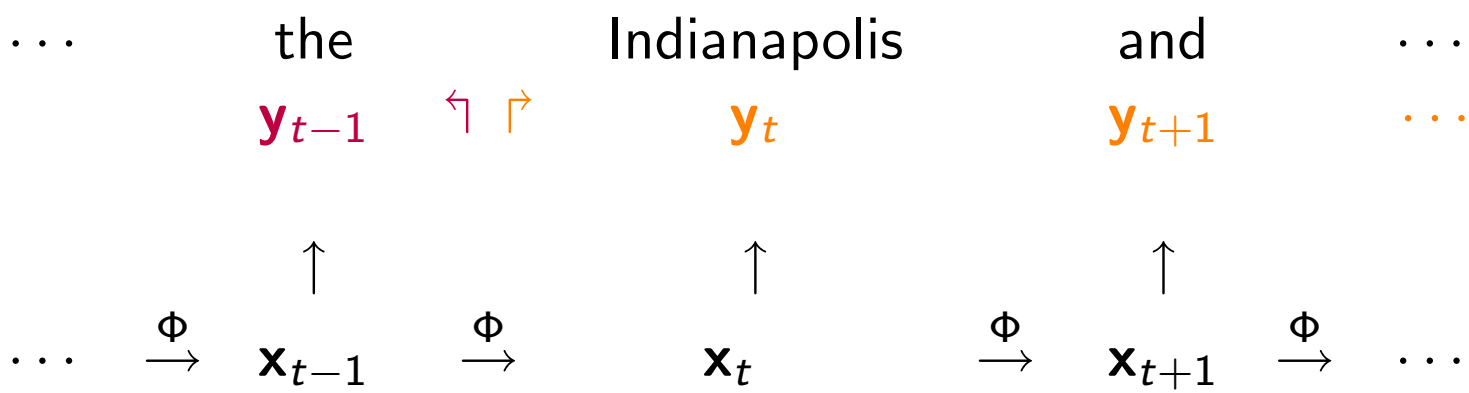
- Extract the current information based on the past and the future



- Let $\mathbf{Y}_t^{(b)} = [\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \cdots]'$

A CCA approach

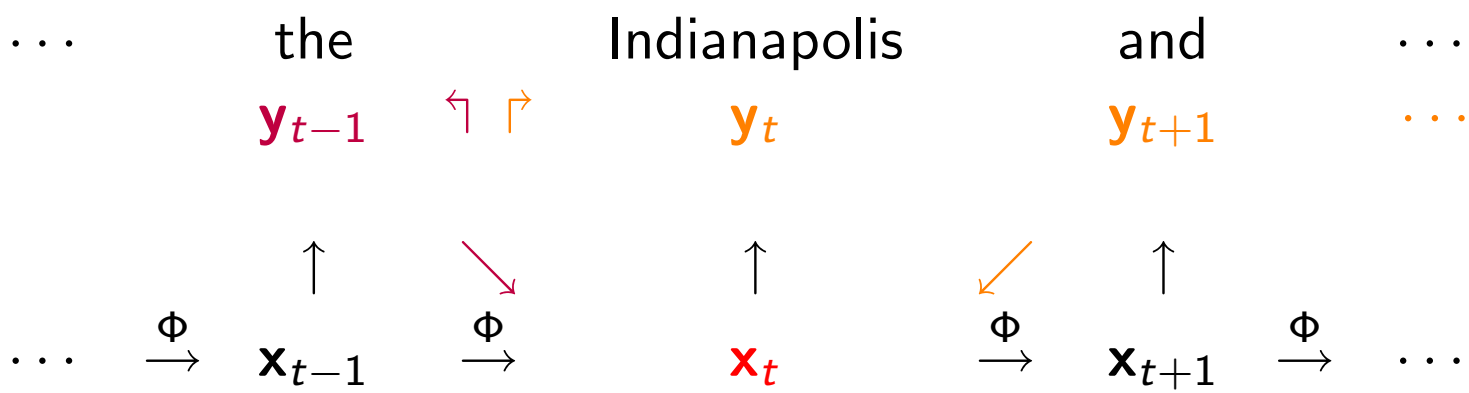
- Extract the current information based on the past and the future



- Let $\mathbf{Y}_t^{(b)} = [\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \cdots]'$
- Let $\mathbf{Y}_t^{(f)} = [\mathbf{y}'_t, \mathbf{y}'_{t+1}, \cdots]'$

A CCA approach

- Extract the current information based on the past and the future



- Let $\mathbf{Y}_t^{(b)} = [\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \cdots]'$
- Let $\mathbf{Y}_t^{(f)} = [\mathbf{y}'_t, \mathbf{y}'_{t+1}, \cdots]'$
- $\mathbf{x}_t = \Sigma^{1/2} V' \mathbf{Y}_t^{(b)}$ where Σ and V are attained from SVDs of the covariances $\mathbb{E}\mathbf{Y}_t^{(b)}(\mathbf{Y}_t^{(b)})'$, $\mathbb{E}\mathbf{Y}_t^{(f)}(\mathbf{Y}_t^{(f)})'$ and $\mathbb{E}\mathbf{Y}_t^{(b)}(\mathbf{Y}_t^{(f)})'$.

Application – The WikipediA Project

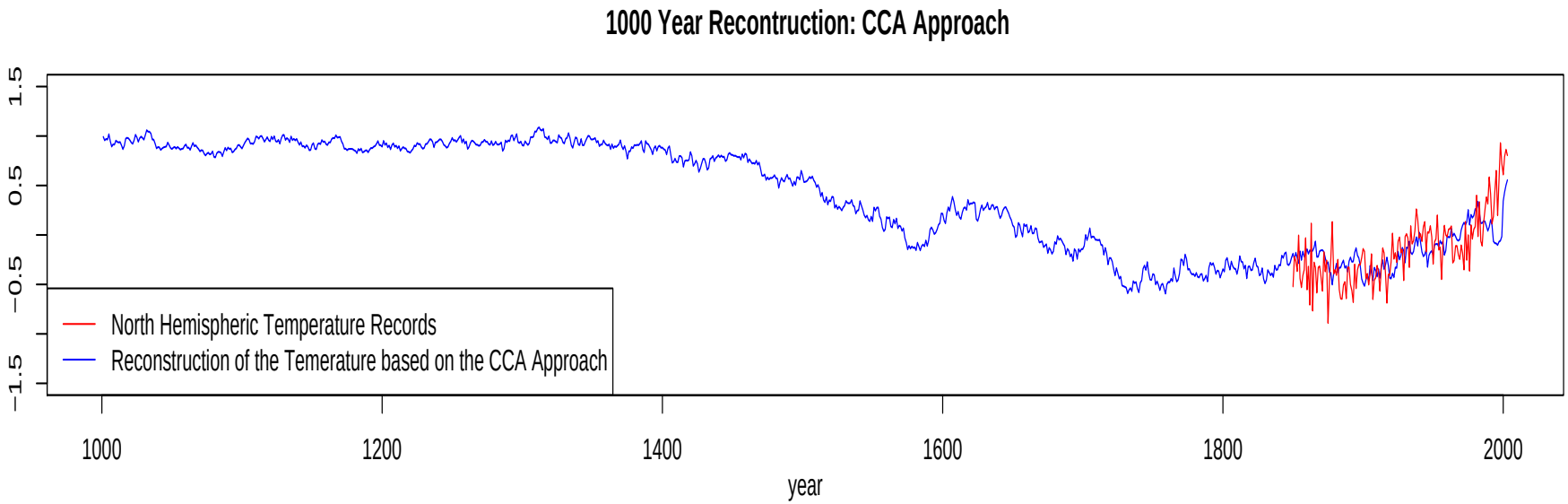
As shown in the table below, the state-based regression gives much better results than the other methods.

Method	Accuracy
Random guessing	50%
Unsupervised state-based	61%
TF-IDF	62%
State-based regression	77%

Other Application

The approach is so simple that it can be widely applied to high dimensional datasets

- Global Warming
- Multiple Learning Skills
- Speaker Identification using video streams



Thank You!