# Person Name Segmentation with Deep Neural Networks

T. Y. S. S. Santosh, Debarshi Kumar Sanyal, Partha Pratim Das

Indian Institute of Technology Kharagpur

MIKE 2019
NIT, Goa

# Overview

# Introduction

- Many applications require organizing personal names in a consistent format.
  - Library catalogs and bibliographies mention the last name first.
  - Common requirement for author metadata at the National Digital Library of India (NDLI).
- Difficult to write a rule-based system due to diversity of names.
- We explore a deep learning-based approach for segmenting person names automatically.

# Related Work

- Existing techniques are of 3 types: (1) rule-based, (2) statistical learning-based, and (3) hybrid.
- Statistical learning techniques either use generative models like HMM or discriminative models like CRFs.
  - HMM is used for address and name segmentation in [1].
  - References [2] and [3] employed HMMs to normalize Australian person names and person names in medical databases respectively.
  - Das et al. [4] used CRF for parsing names in a LinkedIn dataset.
- The choice of the model often depends on the application with no clear winner among them [5].
- Deep learning is very popular nowadays [6]. Recurrent Neural Networks (RNNs) look promising for our problem.

# Problem Definition

- Input sequence $X =< x_1, x_2, \cdots, x_n >$ comprises the components in the name.
- Target sequence $Y =< y_1, y_2, \cdots, y_n >$ comprises the labels of the components.
- Example:
  - Name: `Sharma Ramesh Chandra`
  - Input sequence: $X$ = `<Sharma, Ramesh, Chandra>`
  - Target sequence: $Y$ = `<LN, RN, RN>`
- In practice, we seek $Y^*$ that maximizes the conditional probability $p(Y|X, \Lambda)$ where $\Lambda$ is the set of model parameters:

$$Y^* = \arg\max_{Y} p(Y|X, \Lambda) \tag{1}$$

$$p(Y|X, \Lambda) = p(y_1, \cdots, y_n | x_1, \cdots x_n, \Lambda) \tag{2}$$

# Contributions

- We use RNN-based model to segment person names automatically.
- We evaluate our model on a large corpus of person names from NDLI. It shows an accuracy of 94% while an HMM produces 83.5% accuracy.
- We show visualizations of the learned representations.

# RNN-Based Segmenter

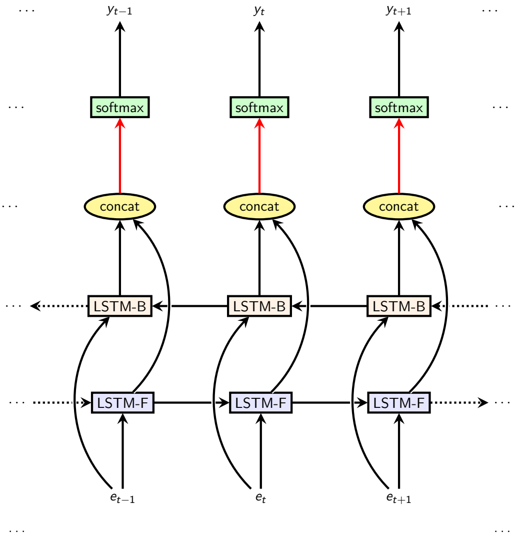



Figure: BiLSTM to segment person names.

# RNN-Based Segmenter

- Two models variants with different the output layers have been designed:
  1. BiLSTM with softmax layer.
  2. BiLSTM with CRF layer.
- Each of the above architectures is again subdivided into three types based on the input:
  1. word level.
  2. character level.
  3. word + character level.

# HMM-Based Segmenter (contd.)

- As an alternative to the deep learning model, a Hidden Markov Model (HMM) has been designed to map name components to states.
  - states $\in$ {START,LN,SFX,RN,END};
  - $n \times n$ *state transition matrix* $\mathbf{A} = [a_{ij}]$;
  - $n \times m$ *emission probability matrix* $\mathbf{B} = [b_{jk}]$ where $b_{jk} = b_j(w_k)$, the probability of emitting symbol $w_k$ in state $j$:
    $a_{ij} = \frac{\text{Number of transitions from state } i \text{ to state } j}{\text{Total number of transitions out of state } i}$
    $b_j(w_k) = \frac{\text{Number of times } w_k \text{ is emitted from state } j}{\text{Total number of symbol-emissions from state } j}$
- The Viterbi algorithm is applied to find the most likely state sequence $Y^*$ that is generated by the input sequence $X$.

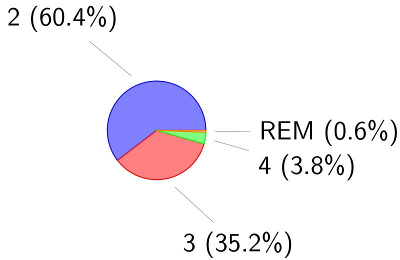
# HMM-Based Segmenter (contd.)

- We use the following 2 smoothing techniques separately.
  1. Laplace Smoothing: we choose a pseudocount $\mu = 1$ and assume that each symbol in $V$ appears at least $\mu$ times so that [UNK] does not get zero probability.
  2. Absolute Discounting:
     - We subtract $\delta > 0$ from the emission probability of each known symbol $w_k$ emitted from state $j$. So, new emission probability of $w_k$ is $b'_j(w_k) = b_j(w_k) - \delta$ in state $j$.
     - The total subtracted probability is divided equally among the symbols not seen in state $j$.
     - Thus, if $T_j$ unique symbols are seen in state $j$ during training, the probability of an unseen symbol to be emitted from state $j$ is $\frac{T_j \delta}{m - T_j}$.
     - We choose $\delta = \frac{1}{T_j + m}$ [1].

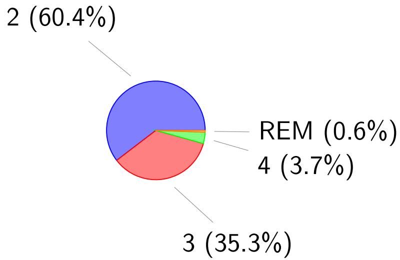
# Dataset

- Our corpus contains author names from IEEE publications indexed in NDLI.
- Names are in `<LN+, SFX?, RN+>` format.
- We remove all separating commas and augment the dataset by circular right-shifting each name so that there are `<RN+, LN+, SFX?>` names, too. Otherwise, the segmenter will only learn to output `<LN+, SFX?, RN+>`.
- Corpus is divided in 80 : 20 ratio into training and test subsets.
    - Training subset holds 1.3 million author names.
    - Test subset holds 0.34 million author names.

# Dataset (contd.)




(a) Training corpus. `REM` comprises names of lengths 1,5,6,7,8,9.




(b) Test corpus. `REM` comprises names of lengths 1,5,6,7,8,9.

Figure: Distribution of the number of components in an author name in the corpus.

# Results

| Model | Vocabulary size (#words) | Accuracy (%) |
|---|---|---|
| WordEmb-BiLSTM-SoftMax | 30K | 90.05 |
| CharacterEmb-BiLSTM-SoftMax | X | 93.78 |
| (Word+Char)Emb-BiLSTM-SoftMax | 30K | 92.64 |
| WordEmb-BiLSTM-CRF | 30K | 91.85 |
| **CharacterEmb-BiLSTM-CRF** | X | **93.97** |
| (Word+Char)Emb-BiLSTM-CRF | 30K | 93.09 |

Table: Performance of deep learning-based segmenters.

# Results (HMM)

| Vocabulary size (#words) | Smoothing function | Accuracy (%) |
|---|---|---|
| 30K | Laplace | 83.5 |
| 30K | Absolute discounting | 81.98 |

Table: Performance of HMM-based segmenter.

# Visualization of Learned Representations

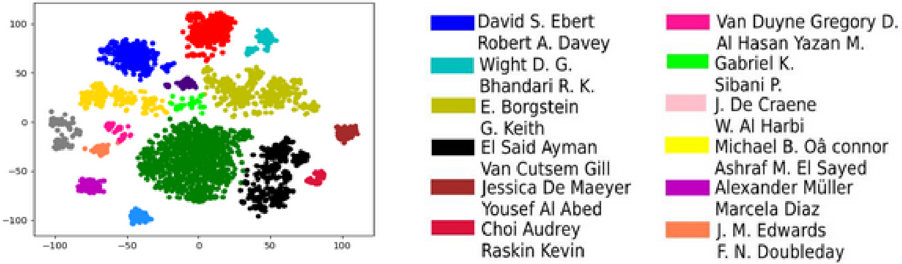



Figure: Name embeddings clustered with DBSCAN

# Summary and Future Work

- We presented a novel deep learning-based name segmentation technique.
- The character BiLSTM with CRF achieved an accuracy of 94%.
- BiLSTM with CRF outperformed BiLSTM with softmax and both vastly outperformed HMM.
- Character model was found superior to word or combination models for the name segmentation task.
- Our results set a baseline for more complex name segmentation techniques.
- We would also explore if active learning can increase the accuracy further.

# References I


- V. Borkar, K. Deshmukh, and S. Sarawagi, "Automatic segmentation of text into structured records," in *ACM SIGMOD Record*, vol. 30. ACM, 2001, pp. 175–186.
- T. Churches, P. Christen, K. Lim, and J. X. Zhu, "Preparation of name and address data for record linkage using hidden markov models," *BMC Medical Informatics and Decision Making*, vol. 2, no. 1, p. 9, 2002.
- R. d. C. B. Gonçalves and S. M. Freire, "Name segmentation using hidden markov models and its application in record linkage," *Cadernos de Saude Publica*, vol. 30, no. 10, pp. 2039–2048, 2014.

# References II


- G. S. Das, X. Li, A. Sun, H. Kardes, and X. Wang, "Person-name parsing for linking user web profiles," in *Proceedings of the 18th International Workshop on Web and Databases*. ACM, 2015, pp. 20–26.
- S. Sarawagi, "Information extraction," *Foundations and Trends in Databases*, vol. 1, no. 3, pp. 261–377, 2008.
- L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e2, 2014.


This work is supported by the *National Digital Library of India* Project sponsored by the Ministry of Human Resource Development, Government of India at IIT Kharagpur.