Optimal Transport-based Alignment of Learned Character Representations for String Similarity
Record Linkage

<table>
<thead>
<tr>
<th>US Patent Assignee Records</th>
<th>Company Name</th>
<th>Location</th>
<th>Patent Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethicon Endo Surgery</td>
<td>Somerville, NJ, US</td>
<td>Pneumatically Actuated Surgical Stapler Head</td>
<td></td>
</tr>
</tbody>
</table>

Coreference and Entity Linking

- Excited for these Grammys! Just a weird opening with Tay Sway.
- T-Swift opens the #Grammys
- Always get goosebumps before the #Grammys!!! T-Swift is on!

Search

MoMA Museum of Modern Art

Disambiguation

Shannon

Claude Shannon
HMS Shannon
Shannon County
Shannon (horse)
Similarity of mention strings informs whether or not they refer to the same entity.
String Similarity for Entity Aliases

Which strings can refer to the same entity?

Design similarity function $f$

- $f(string1, string2)$ high similarity if can refer to the same entity
- $f(string1, string2)$ low similarity if cannot refer to the same entity

HMS Shannon
Shannon
County
Shannon

is-alias
not-alias

Designed to inform coreference decisions
Classic Approaches

Similarity determined by number and type of edits

Music in Chile

Music in China

Chinese Music

# Edits = 2

# Edits = 12

Character edits alone insufficient!
STANCE
Similarity of Transport Aligned Neural Character Encodings

Character Embeddings

Optimal Transport based Alignment

CNN Scoring Function
STANCE
Similarity of Transport Aligned Neural Character Encodings

Character Embeddings

Optimal Transport based Alignment

CNN Scoring Function
Character Representations

Encode with RNN, Measure Pairwise Similarities

\[ h_i = \text{RNN}(x_i, h_{i-1}) \]

\[ h_i = \text{RNN}(x_i, h_{i-1}) \]
Character Representations
Encode with RNN, Measure Pairwise Similarities

Repeated characters may suffer from spurious high similarities
STANCE

Similarity of Transport Aligned Neural Character Encodings

Character Embeddings

Optimal Transport based Alignment

Chinese Music

Music in China

CNN Scoring Function

Chinese Music

Music in China
Optimal Transport-based Alignment

Similarity Matrix

Repeated characters may suffer from spurious high similarities

Low Similarity

High Similarity

Chinese Music

Music in China
Optimal Transport-based Alignment

Each character aligned to closest character(s) in other string

Similarity Matrix

Keep "good" alignments

Remove "bad" alignments

Low Similarity

High Similarity

Chinese Music

Music in China
The amount of transported mass indicates degree of alignment.
Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.
Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.
Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.

Cost of transport inversely proportional to similarity.
Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.

Cost of transport inversely proportional to similarity.

To transport:
\[
\text{mass}(M) = \frac{1}{\text{StringLength}}
\]

To receive:
\[
\text{mass}(M) = \frac{1}{\text{StringLength}}
\]

Characters have fixed amount of mass to transport (or receive).

All characters must transport (or receive) entire mass.
Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.

Cost of transport inversely proportional to similarity.

AlignmentCost(S1,S2) = \[ \sum_{i \in S1} \sum_{j \in S2} T_{i \to j} \text{Cost}(i, j) \]

Characters in S1 (e.g., “Chinese Music”)
Characters in S2 (e.g., “Music in China”)
How much of i is transported to j
Inversely proportional to similarity of i & j
Alignment as Optimal Transport

The amount of transported mass indicates degree of alignment.

Cost of transport inversely proportional to similarity of i & j

Characters in S1 (e.g., “Chinese Music”)

Characters in S2 (e.g., “Music in China”)

How much of i is transported to j

Inversely proportional to similarity of i & j

Find minimum cost alignment between characters of the two strings
Optimal Transport-based Alignment

Minimum cost soft alignment btw characters of the two strings

Transport Matrix \( (T) \)

Optimal Transport-based Alignment

How *aligned* the \textit{i} in Chinese is to the \textit{i} in China

Solved with Sinkhorn’s Algorithm: Efficient and differentiable

Sparsity in transport removes spurious matches

\[ \sum = \text{mass}(i) \]
Optimal Transport-based Alignment

Re-weight similarity by transport matrix
STANCE
Similarity of Transport Aligned Neural Character Encodings

Character Embeddings

Optimal Transport based Alignment

CNN Scoring Function
CNN Scoring Function

Capture patterns of sequential alignment between characters.
Experimental Results

Task 1: Alias Detection

Task 2: Cross Document Coreference

Qualitative Analysis & Ablation Study
Alias Detection

Aliases - Two strings that can refer to the same entity

Given a query string, rank candidate aliases.

Query
- Peace Agreement

Candidates
- Peace Treaty
- Peace Pact
- Lease Agreement
- Peacekeeping Troops

Ranking
- Peace Treaty
- Peace Pact
- Peacekeeping Troops
- Lease Agreement
Datasets

Built 5 datasets for alias detection from open KBs

Wikipedia

Irish music is-alias Irish Folk
Datasets

Built 5 datasets for alias detection from open KBs

Wikipedia  Wikipedia-People

Queen Elizabeth II is-alias
Queen Elizabeth the Second
Datasets

Built 5 datasets for alias detection from open KBs

Wikipedia  Wikipedia-People  Music Artist

Red Hot Chili Peppers is-alias RHCP
Datasets

Built 5 datasets for alias detection from open KBs

Wikipedia  Wikipedia-People  Music Artist

Patent Assignee

The Proctor & Gamble Company is-alias Proctor and Gamble
Datasets

Built 5 datasets for alias detection from open KBs

- Wikipedia
- Wikipedia-People
- Music Artist
- Patent Assignee
- Disease

black water fever is-alias hemolytic malaria
Alias Detection Experiments

Compare STANCE to 8 baseline methods including:

Alignment Methods

• Levenshtein Similarity

• Learned Dynamic Time Warping - LDTW (Cuturi et al. 2017)

Neural Methods

• Deep Conflation Model - DCM (Gan et al. 2017)
Alias Detection Experiments

Compare STANCE to 8 baseline methods including:

Alignment Methods

- Levenshtein Similarity
- Learned Dynamic Time Warping - LDTW (Cuturi et al. 2017)

Neural Methods

- Deep Conflation Model - DCM (Gan et al. 2017)
Alias Detection Experiments

Compare STANCE to 8 baseline methods including:

Alignment Methods

• Levenshtein Similarity

• Learned Dynamic Time Warping - LDTW (Cuturi et al. 2017)

Neural Methods

• Deep Conflation Model - DCM (Gan et al. 2017)
Alias Detection Experiments

Compare STANCE to 8 baseline methods including:

Alignment Methods

• Levenshtein Similarity

• Learned Dynamic Time Warping - LDTW (Cuturi et al. 2017)

Neural Methods

• Deep Conflation Model - DCM (Gan et al. 2017)
Alias Detection - Mean Average Precision (MAP)

![Bar chart showing MAP for different categories including Music Artist, Assignee, Disease, Wikipedia, and Wikipedia-People. Categories are compared using Levenshtein Distance, Learned Dynamic Time Warping, and Deep Conflation Model.](chart.png)
Experimental Results

Task 1: Alias Detection

Task 2: Cross Document Coreference

Qualitative Analysis & Ablation Study
Cross-Document Coreference

Twitter at the Grammy’s Dataset (Dredze et al, 2016)
4577 Mentions, 273 Entities
Cross-Document Coreference

Twitter at the Grammy’s Dataset (Dredze et al, 2016)
4577 Mentions, 273 Entities

Excited for these Grammys! Just a weird opening with Tay Sway.

T-Swift opens the #Grammys

Always get goosebumps before the #Grammys!!! Taylor Swift is on!

Taylor, what happened, this is madness. #grammys

LL Cool J has swag for days. No better person to host the #Grammys!

El-El Cool John. #Grammy

LL Cool James just mispronounced @edsheeran's name AGAIN at the #Grammys!
Cross-Document Coreference

Our approach

Average-Linkage Hierarchical Agglomerative Clustering

Use pre-trained STANCE model on Wikipedia-People as pairwise similarity function.

Tune threshold to cut tree for predicting entities on dev set
Cross-Document Coreference Performance

<table>
<thead>
<tr>
<th></th>
<th>B3 F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Green et al (2012)</td>
<td>77.2</td>
</tr>
<tr>
<td>Andrews et al (2014)</td>
<td>72.3</td>
</tr>
<tr>
<td>Green et al (2012)</td>
<td>79.7</td>
</tr>
<tr>
<td>Andrews et al (2014)</td>
<td>72.1</td>
</tr>
<tr>
<td>Andrews et al (2014)</td>
<td>72.3</td>
</tr>
<tr>
<td>Andrews et al (2014)</td>
<td>82.5</td>
</tr>
</tbody>
</table>

Baseline Results from Dredze et al (2016)
Cross-Document Coreference

Twitter at the Grammy’s Dataset (Dredze et al, 2016)

Name variation more informative than context
Experimental Results

Task 1: Alias Detection

Task 2: Cross Document Coreference

Qualitative Analysis & Ablation Study
Qualitative Analysis

Query: Boom Microphones

Nearest Neighbors:

**STANCE**
- Boom mike
- Boom mics

**LDTW**
- Open Microphone
- Shotgun Microphone

**DCM**
- Open Microphone
- Condensor Microphone
Qualitative Analysis

Query: RPM

Nearest Neighbors:

<table>
<thead>
<tr>
<th>STANCE</th>
<th>RPM Weekly</th>
<th>Randle Patrick McMurray</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDTW</td>
<td>PPM</td>
<td>RPM Alternative 30</td>
</tr>
<tr>
<td>DCM</td>
<td>RP1</td>
<td>PRM</td>
</tr>
</tbody>
</table>
Impact of Optimal Transport in STANCE

OT component improves results on 4 of 5 datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>w/o OT</th>
<th>STANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Music Artist</td>
<td>0.475</td>
<td>0.597</td>
</tr>
<tr>
<td>Assignee</td>
<td>0.36</td>
<td>0.417</td>
</tr>
<tr>
<td>Disease</td>
<td>0.34</td>
<td>0.416</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>0.538</td>
<td>0.594</td>
</tr>
<tr>
<td>Wikipedia-People</td>
<td>0.594</td>
<td>0.594</td>
</tr>
</tbody>
</table>
Benefit of OT - Noise Reduction

Query: Saath Saath Banayenger Ek Aashi

Non-Alias Candidate: Teen Bahuraaniyaan

Significant number of repeated characters and character bigrams

Similarity Matrix - w/o OT
Benefit of OT - Noise Reduction

Query: Saath Saath Banayenger Ek Aashi

Non-Alias Candidate: Teen Bahuraaniyaan

Significant number of repeated characters and character bigrams

Similarity Matrix - w/o OT
Benefit of OT - Noise Reduction

Query: Saath Saath Banayenger Ek Aashi

Non-Alias Candidate: Teen Bahuraaniyaan

Significant number of repeated characters and character bigrams

Similarity Matrix - w/o OT

Similarity Matrix - STANCE
Summary

Learned String Similarity

STANCE
Similarity of Transport Aligned Neural Character Encodings

New Datasets and Results
Thanks! Questions?

Code: https://github.com/iesl/stance