Measures of Distributional Similarity, Lillian Lee, ACL '99

= similarity between words, represented as co-occurrence distributions/vectors.

It can be *really hard* to read your past work.

OTOH, "Anyone who isn't embarrassed of who they were last year probably isn't learning enough."
– attributed to Alain de Botton

Image credit: a family member who wishes to remain anonymous

(See also Own Your Words, Millie Florence)

Measures of Distributional Similarity, Lillian Lee, ACL ‘99

= similarity between words, represented as co-occurrence distributions/vectors.

It can be *really hard* to read your past work.

OTOH, “Anyone who isn't embarrassed of who they were last year probably isn't learning enough.”
– attributed to Alain de Botton

Image credit: a family member who wishes to remain anonymous

Let’s turn on reviewer mode.

(See also Own Your Words, Millie Florence)

Strengths

Weaknesses

1. Language-modeling objective holds up.
    a. Goal: Improve prediction of P(string$_2$ | string$_1$)
    b. Use distributional representation where similar vectors ≊ similar syntax/semantics

2. Earliest ACL Anthology mention of the *Jensen-Shannon divergence.*

3. Introduced the *skew divergence*: intuitive, more stable approximation of the KLD. Has been used in (see Kimura and Hino ‘21 for refs): image recognition, graph analysis, quantum information theory; an ICML ‘24 paper used reverse s.d. as improved loss function.

4. Maybe it’s wrong/the nostalgia speaking, but I’m still fond of the analysis/experiments/information theory.

Strengths

Weaknesses

1. Language-modeling objective holds up.
   a. Goal: Improve prediction of P(string$_2$ | string$_1$)

      The paper focused on *smoothing* — (the 1990s' sense is) dead

   b. Use distributional representation where similar vectors ≋ similar syntax/semantics

2. Earliest ACL Anthology mention of the *Jensen-Shannon divergence.*

3. Introduced the *skew divergence*: intuitive, more stable approximation of the KLD. Has been used in (see Kimura and Hino '21 for refs): image recognition, graph analysis, quantum information theory; an ICML '24 paper used reverse s.d. as improved loss function.

4. Maybe it's wrong/the nostalgia speaking, but I'm still fond of the analysis/experiments/information theory.

Strengths

Weaknesses

1. Language-modeling objective holds up.
   a. Goal: Improve prediction of P(string) or P(string$_2$ | string$_1$)

      The paper focused on *smoothing* — (the 1990s' sense is) dead

   b. Use distributional representation where similar vectors ≈ similar syntax/semantics

      Based on *sparse* (co-occurrence) distributions — dead (for now).

2. Earliest ACL Anthology mention of the *Jensen-Shannon divergence.*

3. Introduced the *skew divergence*: intuitive, more stable approximation of the KLD. Has been used in (see Kimura and Hino '21 for refs): image recognition, graph analysis, quantum information theory; an ICML '24 paper used reverse s.d. as improved loss function.

4. Maybe it's wrong/the nostalgia speaking, but I'm still fond of the analysis/experiments/information theory.

Strengths

Weaknesses

1. Language-modeling objective holds up.
   a. Goal: Improve prediction of P(string$_2$ | string$_1$) or P(string$_2$ | string$_1$)

      The paper focused on *smoothing* — (the 1990s' ser

   b. Use distributional representation where similar vectors ≈ similar syntax

      Based on *sparse* (co-occurrence) distributions — de

Comments

Sparse-data problem: is it still a thing?

2. Earliest ACL Anthology mention of the *Jensen-Shannon divergence.*

3. Introduced the *skew divergence*: intuitive, more stable approximation of the KLD. Has been used in (see Kimura and Hino '21 for refs): image recognition, graph analysis, quantum information theory; an ICML '24 paper used reverse s.d. as improved loss function.

4. Maybe it's wrong/the nostalgia speaking, but I'm still fond of the analysis/experiments/information theory.

Strengths

Weaknesses

1. Language-modeling objective holds up.
   a. Goal: Improve prediction of P(string$_2$ | string$_1$)

      The paper focused on *smoothing* — (the 1990s' ser

   b. Use distributional representation where similar vectors ≈ similar synta

      Based on *sparse* (co-occurrence) distributions — de

2. Earliest ACL Anthology mention of the *Jensen-Shannon divergence.*

   … not counting Lee & F. Pereira, ACL '99. Fernando made the connection.

3. Introduced the *skew divergence*: intuitive, more stable approximation of the KLD. Has been used in (see Kimura and Hino '21 for refs): image recognition, graph analysis, quantum information theory; an ICML '24 paper used reverse s.d. as improved loss function.

4. Maybe it's wrong/the nostalgia speaking, but I'm still fond of the analysis/experiments/information theory.

Comments

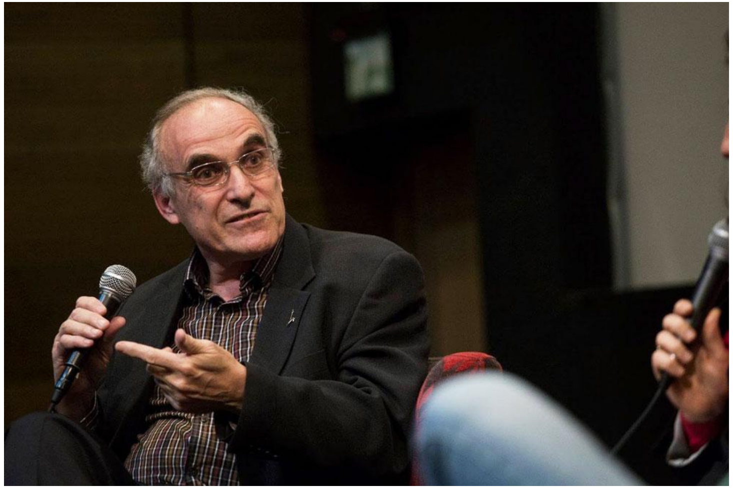
Sparse-data problem: is it still a thing?

Knowledge of or educated guess at author identity

Image credits: Marcos Borga, https://images.trustinnews.pt/uploads/sites/5/2019/12/1581009?Fernando-Pereira-Google.jpeg; https://2021.emnlp.org/keynotes; Miriam Alster, https://d2r55xnwy6nx47.cloudfront.net/uploads/2017/09/Naftali-Tishby_LR.jpg

Fernando Pereira

Ido Dagan



Naftali Tishby (1952-2021)

Please consider this a joint award!

PTL93+DPL94+DLP97+DLP99+LP99+L99+L01

With much debt to their earlier independent work, PT92 and D-Marcus-Markovitch93.

Thanks also to…

- **Stuart Shieber**, my Ph.D. advisor, who hosted and paid me (!) to work on this paper at Harvard after my 1st summer as an assistant professor, but would never claim coauthorship if he didn't believe he had contributed at least 33%.
- **Matt Post** and **Marcel Bollmann**, for some ACL-anthology historical-data wrangling for this talk.

- **The nominating committee and whoever nominated the paper.** I can only hope to pay forward your selfless actions. Thank you very much.