Toward Socially-Infused Information Extraction: Embedding Authors, Mentions, and Entities

Yi Yang*, Ming-Wei Chang#, and Jacob Eisenstein*

* Georgia Tech, # Microsoft Research

Entity Linking on Microblogs

Challenge: insufficient local context

Solution: *homophily* theory

- Neighbors have similar properties.

Entity homophily: socially linked individuals share similar interests in entities.

Testing Entity Homophily

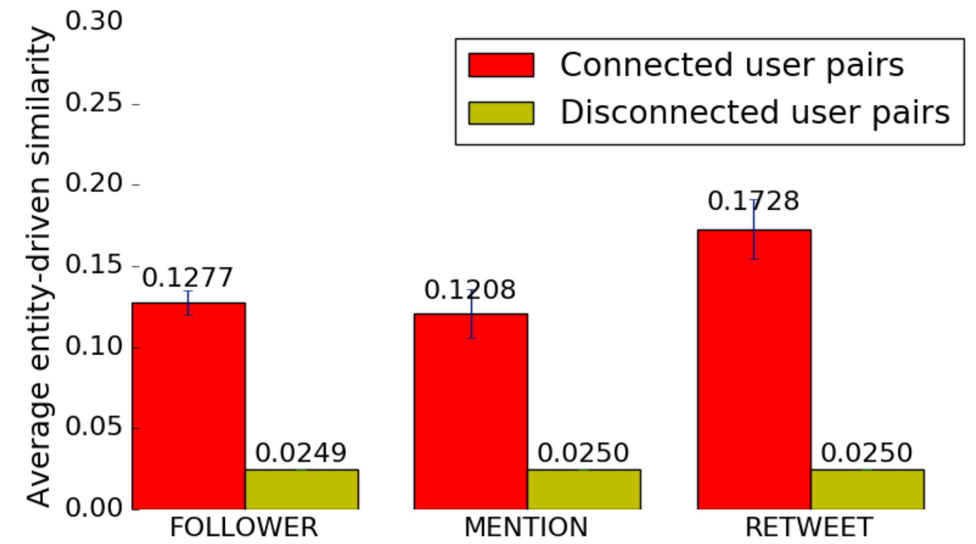
Data: NEEL training data (Cano et al., 2014)

Network	# Author	# Relation
FOLLOWER	1,317	1,604
MENTION	1,317	379
RETWEET	1,317	342

Metric: *entity-driven similarity* between authors

- Cosine similarity between the vectors of entities mentioned by authors (Ritter et al., 2011).

Results (and 90% confidence intervals)



Representations

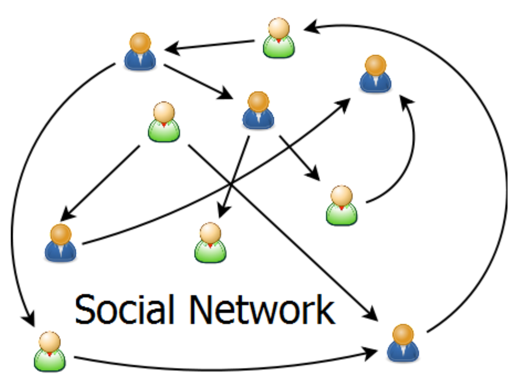
Surface features: $\phi(\mathbf{x}, y_t, t)$

Statistical dense features used by Yang and Chang (2015), extracted from

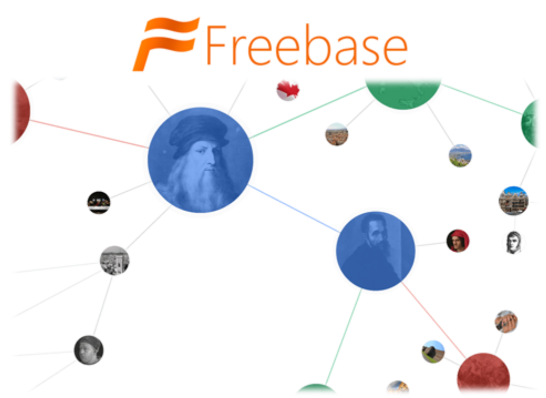
- A named entity recognizer
- An entity type recognizer
- Some statistics of the Wikipedia pages

Distributed representations of authors, mentions, and entities

- Author embeddings (Tang et al., 2015): $\mathbf{v}_u^{(u)}$
  - social connected users are close to each other in the embedding space.



- Mention embeddings (Ling et al., 2015): $\mathbf{v}_t^{(m)}$
  - the average of embeddings of words that the mention contains: $\overrightarrow{\text{Red Sox}} = (\overrightarrow{\text{Red}} + \overrightarrow{\text{Sox}})/2$
- Entity embeddings (Mikolov et al., 2013): $\mathbf{v}_{y_t}^{(e)}$
  - the pre-trained Freebase entity embeddings released by Google



Model

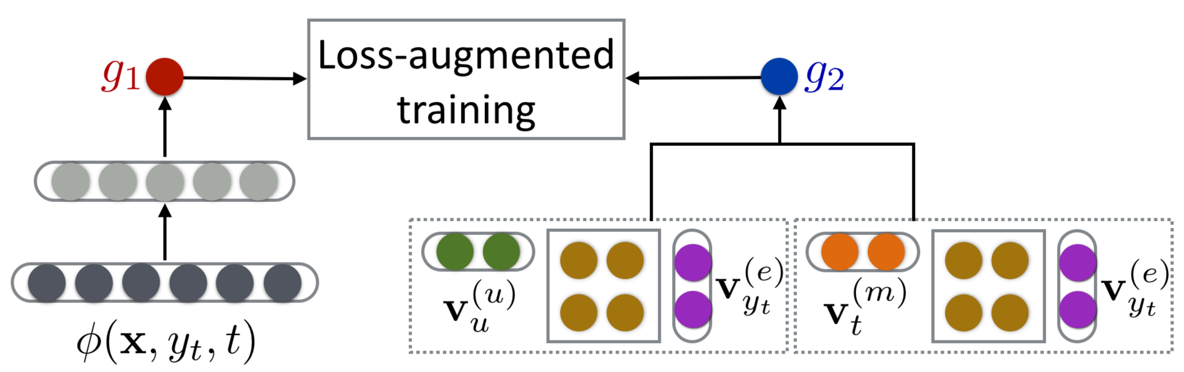
Overview

$$s(\mathbf{x}, \mathbf{y}, u) = g_1(\mathbf{x}, y_t, t) + g_2(\mathbf{x}, y_t, u, t)$$

$$g_1(\mathbf{x}, y_t, t; \Theta_1) = \boldsymbol{\beta}^\top \tanh(\mathbf{W}\phi(\mathbf{x}, y_t, t) + \mathbf{b}) + b$$

$$g_2(\mathbf{x}, y_t, u, t; \Theta_2) = \boxed{\mathbf{v}_u^{(u)}}^\top \mathbf{W}^{(u,e)} \boxed{\mathbf{v}_{y_t}^{(e)}} + \boxed{\mathbf{v}_t^{(m)}}^\top \mathbf{W}^{(m,e)} \boxed{\mathbf{v}_{y_t}^{(e)}}$$

author embedding mention embedding entity embedding



- A multilayer perceptron (MLP) is adopted to model surface features (g_1).
- Two bilinear scoring functions are employed to explicitly leverages the assumptions (g_2):
  - entity homophily
  - semantically related mentions are likely to be linked to similar entities

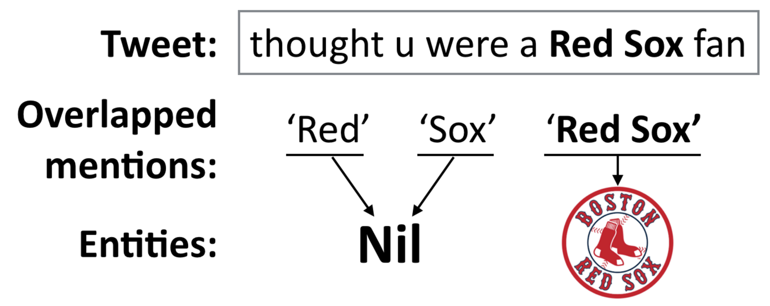
Learning

$$L = \max_{\mathbf{y} \in \mathcal{Y}_\mathbf{x}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u)) - s(\mathbf{x}, \mathbf{y}^*, u)$$

Inference

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}_\mathbf{x}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u))$$

- Non-overlapping structure



In order to link 'Red Sox' to a real entity, 'Red' and 'Sox' should be linked to **Nil**.

Experiments

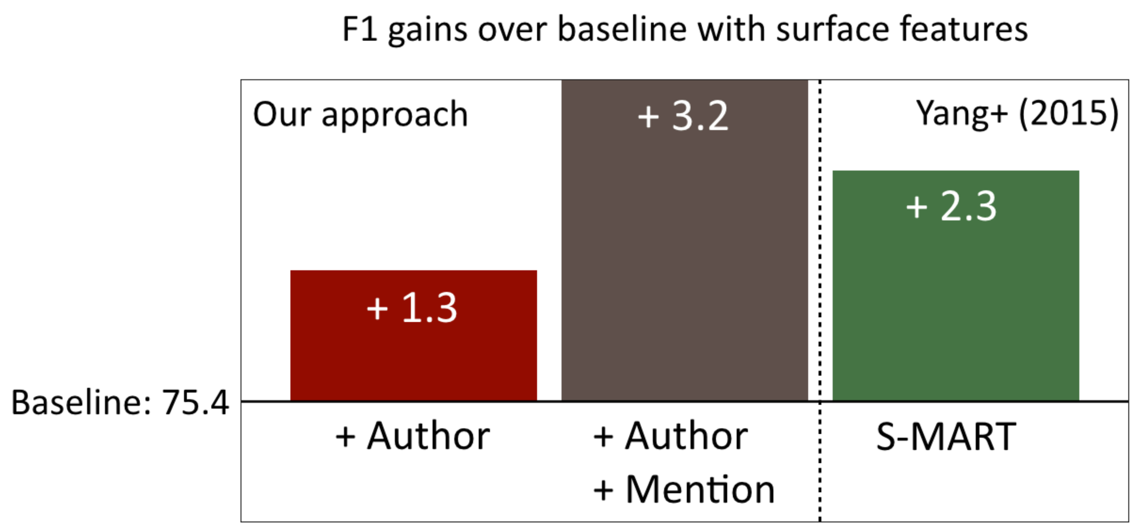
Social network expansion

- We expand the FOLLOWER, MENTION, and RETWEET networks by including nodes that will do the most to densify the author networks.
- The new networks result in better author embeddings.

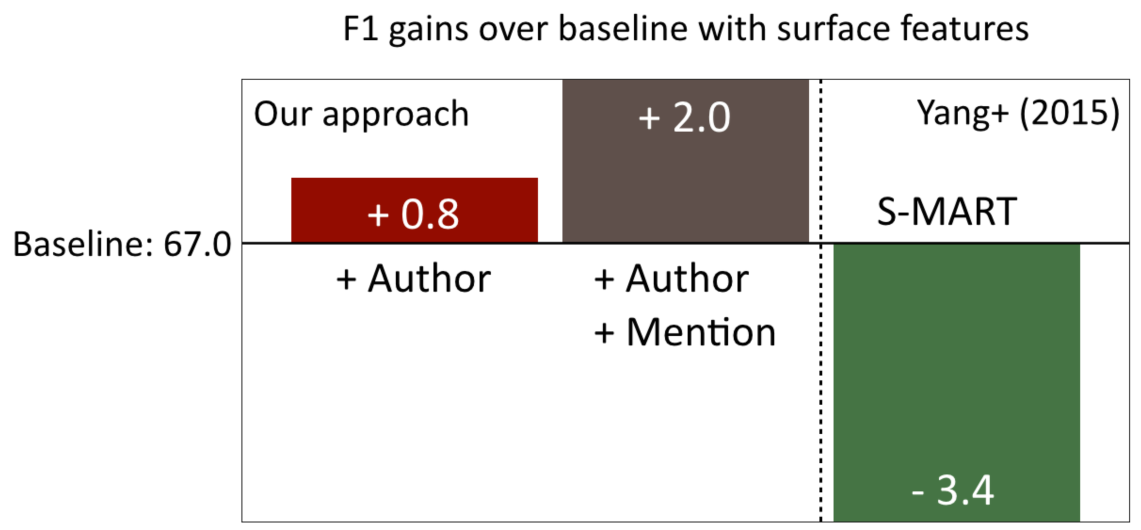
Network	# Author	# Relation
FOLLOWER+	8,772	286,800
MENTION+	6,119	57,045
RETWEET+	7,404	59,313

Results

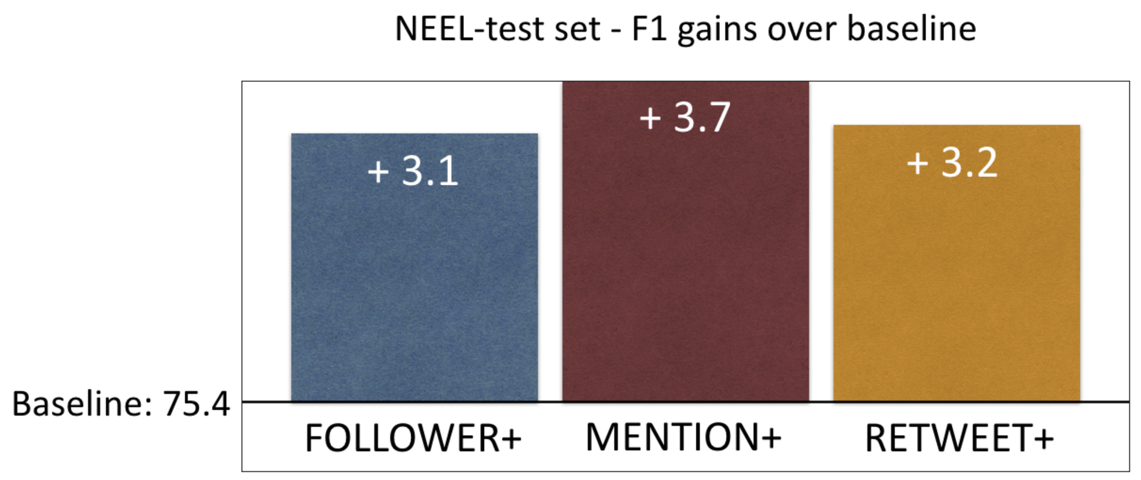
- NEEL-test set (Cano et al., 2014)

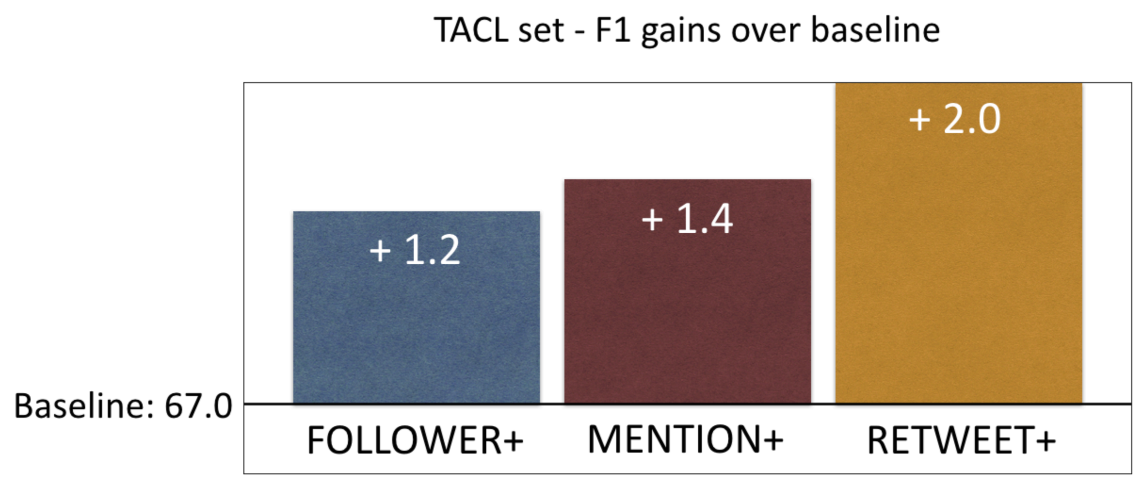


- TACL set (Fang and Chang, 2014)



- Comparison of different social networks





Summary

- We present a novel neural network model for entity linking that exploits distributed representations of users, mentions, and entities.
- Our system leverages social network structures by utilizing entity homophily to improve entity disambiguation.
- Our neural network model is on par with the tree-based model (Yang and Chang 2015) with surface features, but it is much easier to add additional information in the neural network model.