Collective Tweet Wikification based on Semi-supervised Graph Regularization

Hongzhao Huang¹, Yunbo Cao², Xiaojiang Huang², Heng Ji¹, Chin-Yew Lin²

Computer Science Department, Rensselaer Polytechnic Institute¹
Microsoft Research Asia²
At a WH briefing here in Santiago, NSA spox Rhodes came with a litany of pushback on idea WH didn't consult with.
Wikification for Tweets is Important

- Stay up Hawk Fans. We are going through a slump now, but we have to stay positive. Go *Hawks*!
- Go *Gators*!

Also useful for: coreference (Ratinov and Roth, 2012), text classification (Vitale et al., 2012)
Previous Work on Entity Linking/Wikification

- **Prior Popularity** (e.g., Han & Sun, 2011; Meiji et al., 2012)
  - Tend to link a mention to the most probable underlying concept

- **Context Similarity** (e.g., Chen & Ji, 2011; Guo et al., 2013)
  - Tend to link a mention to the concept which are the most overlapping context

- **Topical Coherence** (e.g., Zhang et al., 2011; Han et al., 2011; Shen et al., 2013; Liu et al., 2013)
  - Assume a mention’s referent concept is topically coherent with other mentions with the same context
Wikification for Tweets is Difficult

- Tweets are **short** (up to 140 characters)
  - Lack of rich context to compute context similarity and ensure topical coherence

- Stay up *Hawk Fans*. We are going through a *slump*, but we have to stay positive. Go *Hawks*!

- Go *Gators*!
Wikification for Tweets is Difficult (con’t)

- Lack of **Labeled Data** for Supervised Model
  - Annotation for Wikification is challenging
    - Difficult to select important concept mentions
    - Corresponding concept may not exist
    - Mentions are ambiguous
  - Annotation for short tweets is more challenging
    - Lack of enough background information
- A more appropriate approach for wikification for tweets
  - Handle information shortage problem
  - Demand less labeled data
Deep Analysis

• Mentions and their correct referent concepts tend to share a set of characteristics
  ▪ String similarity (<Chicago, Chicago> and <Facebook, Facebook>)

• Two coreferential mentions should be linked to the same concept
  • User A: At a WH briefing here in Santiago, NSA spox Rhodes…
  • User A: Chinese president is going to visit White House and meet Obama….
Deeper Analysis (cont’)

- Two highly semantically-related mentions are more likely to be linked to two highly semantically-related concepts
  - Stay up *Hawk Fans*. We are going through a *slump*, but we have to stay positive. Go *Hawks*!
  - Go *Gators***!!!
Our Approach: Semi-Supervised Graph Regularization

- The Model (Zhu et.al, 2003):
  \[ Q(Y) = \mu \sum_{i=l+1}^{n} (y_i - y_i^0)^2 + \frac{1}{2} \sum_{i,j} W_{ij}(y_i - y_j)^2. \]

- Perform *collective inference*
  - Identify and disambiguate multiple mentions from multiple messages simultaneously

- Make use of tremendous unlabeled data

---

- Handle the lack of labeled data problem
Relational Graph

- Each pair of mention m and concept c as a node

An example of the relational graph

- m is linkable, and c is the correct concept, <m, c> should be assigned label 1, otherwise 0
Relevant Mention Detection: Meta Path

- A meta-path is a path defined over a network and composed of a sequence of relations between different object types (Sun et al., 2011)
  - Each meta path represent a semantic relation
- Meta paths between mention and mention
  - M-T-M
  - M-T-U-T-M-M
  - M-T-H-T-M
  - M-T-U-T-M-T-H-T-M
  - M-T-H-T-M-T-U-T-M

M: mention, T: tweet, U: user, H: hashtag
Relational Graph Construction

- **Local Compatibility**
  - Mention Features (e.g., idf, keyphraseness)
  - Concept Features (e.g., # of incoming/outgoing links)
  - **Mention + Concept** Features (prior popularity, tf)
  - Context Features (capitalization, tf-idf)
Relational Graph Construction (con’t)

- Coreference
  - At least one meta path exists between two similar mentions
Relational Graph Construction (con’t)

- Semantic Relatedness (SR)
  - SR between two mentions: meta path
  - SR between two concepts: link structure in Wikipedia (Milne and Witten, 2008)
    \[
    SR(c_i, c_j) = 1 - \frac{\log \max(|C_i|, |C_j|) - \log |C_i \cap C_j|}{\log(|C|) - \log \min(|C_i|, |C_j|)}
    \]
  - Linear Combination of the three graphs
Data and Scoring Metric

- **Data**
  - A public data set includes 502 messages from 28 users (Meiji et al., 2012)
  - A Wikipedia dump on May 3, 2013

- **Scoring Metric**
  - Standard precision, recall and F1
  - A pair of mention and concept is judged as correct
    - Mention is linkable
    - Concept is the correct referent concept
Models for Comparison

- TagMe: an unsupervised model based on prior popularity and semantic relatedness of a single message (Ferragina and Scaiella, 2010)
- Meij: the state-of-the-art supervised approach based on the random forest model (Meij et al., 2012)
- SSRegu: our proposed semi-supervised graph regularization model
Overall Performance

- Meij: use 100% labeled data
- SSRegu: use 50% labeled data

Collective inference over multiple mentions and multiple tweets is effective

5% absolute F1 gain over the state-of-the-art supervised models.
Our model with 30% labeled data achieves similar performance with the state-of-the-art supervised model.
Some Examples

- The supervised baseline
  - bucks  Bucks County, Pennsylvania | 0.09
  - uconn  University of Connecticut | 0.19

- Our Approach
  - bucks  *Milwaukee Bucks* | 0.45  Bucks County, Pennsylvania | 0.12
  - uconn  *Connecticut Huskies* | 0.24  University of Connecticut | 0.21  Connecticut Huskies men's basketball | 0.13
Related Work

- Collective approaches for Wikification/Entity Linking (Cucerzan, 2007; Milne and Witten, 2008b; Kulkarni et al., 2009; Pennacchiotti and Pantel, 2009; Ferragina and Scaiella, 2010; Fernandez et al., 2010; Radford et al., 2010; Cucerzan, 2011; Guo et al., 2011; Han et al., 2011; Ratinov et al., 2011; Chen and Ji, 2011; Kozareva et al., 2011; Cassidy et al., 2012; Shen et al., 2013; Liu et al., 2013)
  - We propose a novel graph representation with fine-grained relations
- End-to-End Wikification/Entity Linking (Meiji et al., 2012; Guo et al., 2013)
  - Our method is over multiple tweets and semi-supervised
- Graph-based semi-supervised learning (Zhu et al., 2003; Smola and Kondor, 2003; Zhou et al., 2004; Talukdar and Crammer, 2009)
Conclusions and Future Work

- Proposed a semi-supervised graph regularization model for tweet wikification
  - Collective inference over multiple mentions from multiple tweets
  - Dramatically save annotation costs

- Future Work
  - Learning semantic entity representation to improve graph construction
Thank You!