Weakly Supervised User Profile Extraction from Twitter

Jiwei Li[1], **Alan Ritter**[2], **Eduard Hovy**[1]
[1]Language Technology Institute, [2]Machine Learning Department
Carnegie Mellon University, Pittsburgh, PA 15213, USA
bdlijiwei@gmail.com, rittera@cs.cmu.edu, ehovy@andrew.cmu.edu

Abstract

While user attribute extraction on social media has received considerable attention, existing approaches, mostly supervised, encounter great difficulty in obtaining gold standard data and are therefore limited to predicting unary predicates (e.g., gender). In this paper, we present a weakly-supervised approach to user profile extraction from Twitter. Users' profiles from social media websites such as Facebook or Google Plus are used as a distant source of supervision for extraction of their attributes from user-generated text. In addition to traditional linguistic features used in distant supervision for information extraction, our approach also takes into account network information, a unique opportunity offered by social media. We test our algorithm on three attribute domains: *spouse*, *education* and *job*; experimental results demonstrate our approach is able to make accurate predictions for users' attributes based on their tweets.[1]

1 Introduction

The overwhelming popularity of online social media creates an opportunity to display given aspects of oneself. Users' profile information in social networking websites such as Facebook[2] or Google Plus[3] provides a rich repository personal information in a structured data format, making it amenable to automatic processing. This includes, for example, users' jobs and education, and provides a useful source of information for applications such as search[4], friend recommendation, online advertising, computational social science and more.

@[shanenicholson] has taken all the kids today so I can go shopping-CHILD FREE! #iloveyoushano #iloveyoucreditcard Tamworth promo day with my handsome classy husband @[shanenicholson]
Spouse: shanenicholson
I got accepted to be part of the UofM engineering safety pilot program in [FSU] Here in class. (@ [Florida State University] - Williams Building) Don't worry , guys ! Our beloved [FSU] will always continue to rise " to the top !
Education: Florida State University (FSU)
first day of work at [HuffPo], a sports bar woo come visit me yo.. start to think we should just add a couple desks to the [HuffPo] newsroom for Business Insider writers just back from [HuffPo], what a hell !
Job: HuffPo

Table 1: Examples of Twitter message clues for user profile inference.

Although profiles exist in an easy-to-use, structured data format, they are often sparsely populated; users rarely fully complete their online profiles. Additionally, some social networking services such as Twitter don't support this type of structured profile data. It is therefore difficult to obtain a reasonably comprehensive profile of a user, or a reasonably complete facet of information (say, education level) for a class of users. While many users do not explicitly list all their personal information in their online profile, their user generated content often contains strong evidence to suggest many types of user attributes, for example education, spouse, and employment (See Table 1). Can one use such information to infer more details? In particular, can one exploit indirect clues from an unstructured data source like Twitter to obtain rich, structured user profiles?

In this paper we demonstrate that it is feasible to automatically extract Facebook-style pro-

[1]Both code and data are available at http://aclweb.org/aclwiki/index.php?title=Profile_data
[2]https://www.facebook.com/
[3]https://plus.google.com/
[4]https://www.facebook.com/about/graphsearch

files directly from users' tweets, thus making user profile data available in a structured format for upstream applications. We view user profile inference as a structured prediction task where both text and network information are incorporated. Concretely, we cast user profile prediction as binary relation extraction (Brin, 1999), e.g., SPOUSE(User$_i$, User$_j$), EDUCATION(User$_i$, Entity$_j$) and EMPLOYER(User$_i$, Entity$_j$). Inspired by the concept of distant supervision, we collect training tweets by matching attribute ground truth from an outside "knowledge base" such as Facebook or Google Plus.

One contribution of the work presented here is the creation of the first large-scale dataset on three general Twitter user profile domains (i.e., EDUCATION, JOB, SPOUSE). Experiments demonstrate that by simultaneously harnessing both text features and network information, our approach is able to make accurate user profile predictions. We are optimistic that our approach can easily be applied to further user attributes such as HOBBIES and INTERESTS (MOVIES, BOOKS, SPORTS or STARS), RELIGION, HOMETOWN, LIVING LOCATION, FAMILY MEMBERS and so on, where training data can be obtained by matching ground truth retrieved from multiple types of online social media such as Facebook, Google Plus, or LinkedIn. Our contributions are as follows:

- We cast user profile prediction as an information extraction task.
- We present a large-scale dataset for this task gathered from various structured and unstructured social media sources.
- We demonstrate the benefit of jointly reasoning about users' social network structure when extracting their profiles from text.
- We experimentally demonstrate the effectiveness of our approach on 3 relations: SPOUSE, JOB and EDUCATION.

The remainder of this paper is organized as follows: We summarize related work in Section 2. The creation of our dataset is described in Section 3. The details of our model are presented in Section 4. We present experimental results in Section 5 and conclude in Section 6.

2 Related Work

While user profile inference from social media has received considerable attention (Al Zamal et al., 2012; Rao and Yarowsky, 2010; Rao et al., 2010; Rao et al., 2011), most previous work has treated this as a classification task where the goal is to predict unary predicates describing attributes of the user. Examples include gender (Ciot et al., 2013; Liu and Ruths, 2013; Liu et al., 2012), age (Rao et al., 2010), or political polarity (Pennacchiotti and Popescu, 2011; Conover et al., 2011).

A significant challenge that has limited previous efforts in this area is the lack of available training data. For example, researchers obtain training data by employing workers from Amazon Mechanical Turk to manually identify users' gender from profile pictures (Ciot et al., 2013). This approach is appropriate for attributes such as *gender* with a small numbers of possible values (e.g., *male* or *female*), for which the values can be directly identified. However for attributes such as *spouse* or *education* there are many possible values, making it impossible to manually search for gold standard answers within a large number of tweets which may or may not contain sufficient evidence.

Also related is the Twitter user timeline extraction algorithm of Li and Cardie (2013). This work is not focused on user attribute extraction, however.

Distant Supervision Distant supervision, also known as weak supervision, is a method for learning to extract relations from text using ground truth from an existing database as a source of supervision. Rather than relying on mention-level annotations, which are expensive and time consuming to generate, distant supervision leverages readily available structured data sources as a weak source of supervision for relation extraction from related text corpora (Craven et al., 1999). For example, suppose $r(e_1, e_2) = IsIn(Paris, France)$ is a ground tuple in the database and s ="Paris is the capital of France" contains synonyms for both "Paris" and "France", then we assume that s may express the fact $r(e_1, e_2)$ in some way and can be used as positive training examples. In addition to the wide use in text entity relation extraction (Mintz et al., 2009; Ritter et al., 2013; Hoffmann et al., 2011; Surdeanu et al., 2012; Takamatsu et al., 2012), distant supervision has been applied to multiple

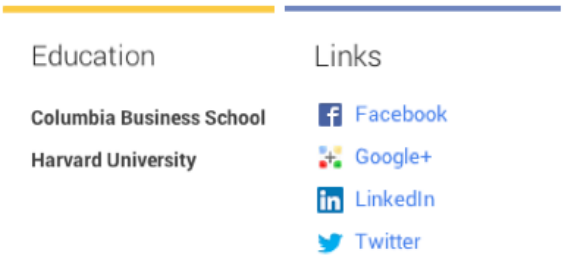

Figure 1: Illustration of Goolge Plus "knowledge base".

Figure 2: Example of fetching tweets containing entity **USC** mention from **Miranda Cosgrove** (an American actress and singer-songwriter)'s twitter stream.

fields such as protein relation extraction (Craven et al., 1999; Ravikumar et al., 2012), event extraction from Twitter (Benson et al., 2011), sentiment analysis (Go et al., 2009) and Wikipedia infobox generation (Wu and Weld, 2007).

Homophily Online social media offers a rich source of network information. McPherson et al. (2001) discovered that people sharing more attributes such as background or hobby have a higher chance of becoming friends in social media. This property, known as HOMOPHILY (summarized by the proverb "birds of a feather flock together") (Al Zamal et al., 2012) has been widely applied to community detection (Yang and Leskovec, 2013) and friend recommendation (Guy et al., 2010) on social media. In the user attribute extraction literature, researchers have considered neighborhood context to boost inference accuracy (Pennacchiotti and Popescu, 2011; Al Zamal et al., 2012), where information about the degree of their connectivity to their pre-labeled users is included in the feature vectors. A related algorithm by Mislove et al. (2010) crawled Facebook profiles of 4,000 Rice University students and alumni and inferred attributes such as major and year of matriculation purely based on network information. Mislove's work does not consider the users' text stream, however. As we demonstrate below, relying solely on network information is not enough to enable inference about attributes.

3 Dataset Creation

We now describe the generation of our distantly supervised training dataset in detail. We make use of Google Plus and Freebase to obtain ground facts and extract positive/negative bags of postings from users' twitter streams according to the ground facts.

Education/Job We first used the Google Plus API[5] (shown in Figure 1) to obtain a seed set of users whose profiles contain both their education/job status and a link to their twitter account.[6] Then, we fetched tweets containing the mention of the education/job entity from each correspondent user's twitter stream using Twitter's search API[7] (shown in Figure 2) and used them to construct positive bags of tweets expressing the associated attribute, namely EDUCATION(User$_i$, Entity$_j$), or EMPLOYER(User$_i$, Entity$_j$). The Freebase API[8] is employed for alias recognition, to match terms such as "Harvard University", "Harvard", "Harvard U" to a single The remainder of each corresponding user's entire Twitter feed is used as negative training data.[9]

We expanded our dataset from the seed users according to network information provided by Google Plus and Twitter. Concretely, we crawled circle information of users in the seed set from both their Twitter and Google Plus accounts and performed a matching to pick out shared users between one's Twitter follower list and Google Plus Circle. This process assures friend identity and avoids the problem of name ambiguity when matching accounts across websites. Among candidate users, those who explicitly display Job or Education information on Google Plus are preserved. We then gathered positive and negative data as described above.

Dataset statistics are presented in Table 2. Our

[5] https://developers.google.com/+/api/

[6] An unambiguous twitter account link is needed here because of the common phenomenon of name duplication.

[7] https://twitter.com/search

[8] http://wiki.freebase.com/wiki/Freebase_API

[9] Due to Twitter user timeline limit, we crawled at most 3200 tweets for each user.

education dataset contains 7,208 users, 6,295 of which are connected to others in the network. The positive training set for the EDUCATION is comprised of 134,060 tweets.

Spouse Facebook is the only type of social media where spouse information is commonly displayed. However, only a tiny amount of individual information is publicly accessible from Facebook Graph API[10]. To obtain ground truth for the spouse relation at large scale, we turned to Freebase[11], a large, open-domain database, and gathered instances of the /PEOPLE/PERSON/SPOUSE relation. Positive/negative training tweets are obtained in the same way as was previously described for EDUCATION and JOB. It is worth noting that our Spouse dataset is not perfect, as individuals retrieved from Freebase are mostly celebrities, and thus it's not clear whether this group of people are representative of the general population.

SPOUSE is an exception to the "homophily" effect. But it exhibits another unique property, known as, REFLEXIVITY: fact $IsSpouseOf(e_1, e_2)$ and $IsSpouseOf(e_2, e_1)$ will hold or not hold at the same time. Given training data expressing the tuple $IsSpouseOf(e_1, e_2)$ from user e_1's twitter stream, we also gather user e_2's tweet collection, and fetch tweets with the mention of e_1. We augment negative training data from e_2 as in the case of Education and Job. Our Spouse dataset contains 1,636 users, where there are 554 couples (1108 users). Note that the number of positive entities (3,121) is greater than the number of users as (1) one user can have multiple spouses at different periods of time (2) multiple entities may point to the same individual, e.g., BarackObama, Barack Obama and Barack.

4 Model

We now describe our approach to predicting user profile attributes.

4.1 Notation

Message X: Each user $i \in [1, I]$ is associated with his Twitter ID and his tweet corpus X_i. X_i is comprised of a collection of tweets $X_i = \{x_{i,j}\}_{j=1}^{j=N_i}$, where N_i denotes the number of tweets user i published.

[10] https://developers.facebook.com/docs/graph-api/

[11] http://www.freebase.com/

	Education	Job	Spouse
#Users	7,208	1,806	1,636
#Users Connected	6,295	1,407	1,108
#Edges	11,167	3,565	554
#Pos Entities	451	380	3121
#Pos Tweets	124,801	65,031	135,466
#Aver Pos Tweets per User	17.3	36.6	82.8
#Neg Entity	6,987,186	4,405,530	8,840,722
#Neg Tweets	16,150,600	10,687,403	12,872,695

Table 2: Statistics for our Dataset

Tweet Collection L_i^e: L_i^e denotes the collection of postings containing the mention of entity e from user i. $L_i^e \subset X_i$.

Entity attribute indicator $z_{i,e}^k$ and $z_{i,x}^k$: For each entity $e \in X_i$, there is a boolean variable $z_{i,e}^k$, indicating whether entity e expresses attribute k of user i. Each posting $x \in L_i^e$ is associated with attribute indicator $z_{i,x}^k$ indicating whether posting x expresses attribute k of user i. $z_{i,e}^k$ and $z_{i,x}^k$ are observed during training and latent during testing.

Neighbor set F_i^k: F_i^k denotes the neighbor set of user i. For Education $(k = 0)$ and Job $(k = 1)$, F_i^k denotes the group of users within the network that are in friend relation with user i. For Spouse attribute, F_i^k denote current user's spouse.

4.2 Model

The distant supervision assumes that if entity e corresponds to an attribute for user i, at least one posting from user i's Twitter stream containing a mention of e might express that attribute. For user-level attribute prediction, we adopt the following two strategies:

(1) GLOBAL directly makes aggregate (entity) level prediction for $z_{i,e}^k$, where features for all tweets from L_i^e are aggregated to one vector for training and testing, following Mintz et al. (2009).

(2) LOCAL makes local tweet-level predictions for each tweet $z_{i,x}^e$, $x \in L_i^k$ in the first place, making the stronger assumption that all mentions of an entity in the users' profile are expressing the associated attribute. An aggregate-level decision $z_{i,e}^k$ is then made from the deterministic OR operators.

$$z_{i,x}^e = \begin{cases} 1 & \exists x \in L_i^e, \mathbf{s.t.} z_{i,x}^k = 1 \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

The rest of this paper describes GLOBAL in detail. The model and parameters with LOCAL are identical to those in GLOBAL except that LOCAL

encode a tweet-level feature vector rather than an aggregate one. They are therefore excluded for brevity. For each attribute k, we use a model that factorizes the joint distribution as product of two distributions that separately characterize text features and network information as follows:

$$\Psi(z_{i,e}^k, X_i, F_i^k : \Theta) \propto \Psi_{text}(z_{i,e}^k, X_i)\Psi_{Neigh}(z_{i,e}^k, F_i^k) \quad (2)$$

Text Factor We use $\Psi^{text}(z_e^k, X_i)$ to capture the text related features which offer attribute clues:

$$\Psi_{text}(z_e^k, , X_i) = \exp[(\Theta_{text}^k)^T \cdot \psi_{text}(z_{i,e}^k, X_i)] \quad (3)$$

The feature vector $\psi_{text}(z_{i,e}^k, X_i)$ encodes the following standard general features:

- Entity-level: whether begins with capital letter, length of entity.
- Token-level: for each token $t \in e$, word identity, word shape, part of speech tags, name entity tags.
- Conjunctive features for a window of k (k=1,2) words and part of speech tags.
- Tweet-level: All tokens in the correspondent tweet.

In addition to general features, we employ attribute-specific features, such as whether the entity matches a bag of words observed in the list of universities, colleges and high schools for *Education* attribute, whether it matches terms in a list of companies for *Job* attribute[12]. Lists of universities and companies are taken from knowledge base NELL[13].

Neighbor Factor For Job and Education, we bias friends to have a larger possibility to share the same attribute. $\Psi_{Neigh}(z_{i,e}^k, F_i^k)$ captures such influence from friends within the network:

$$\Psi_{Neigh}(z_{i,e}^k, F_i^k) = \prod_{j \in F_i^k} \Phi_{Neigh}(z_e^k, X_j)$$
$$\Phi_{Neigh}(z_{i,e}^k, X_j) = \exp[(\Theta_{Neigh}^k)^T \cdot \psi_{Neigh}(z_{i,e}^k, X_j)] \quad (4)$$

Features we explore include the whether entity e is also the correspondent attribute with neighbor user j, i.e., $\mathbf{I}(z_{j,k}^e = 0)$ and $\mathbf{I}(z_{j,k}^e = 1)$.

For Spouse, we set $F_i^{spouse} = \{e\}$ and the neighbor factor can be rewritten as:

$$\Psi_{Neigh}(z_{i,e}^k, X_j) = \Psi_{Neigh}(C_i, X_e) \quad (5)$$

[12]Freebase is employed for alias recognition.

[13]http://rtw.ml.cmu.edu/rtw/kbbrowser/

Input: Tweet Collection $\{X_i\}$, Neighbor set $\{F_i^k\}$
Initialization:
• for each user i:
  for each candidate entity $e \in X_i$
    $z_{i,e}^k = \text{argmax}_{z'} \Psi(z', X_i)$ from text features
End Initialization

while not convergence:
• for each user i:
  update attribute values for $j \in F_i^k$
  for each candidate entity $e \in X_i$
    $z_{i,e}^k = \text{argmax}_{z'} \Psi(z', X_i, F_i^k)$
end while:

Figure 3: Inference for NEIGH-LATENT setting.

It characterizes whether current user C_i to be the spouse of user e (if e corresponds to a Twitter user). We expect clues about whether C_i being entity e's spouse from e's Twitter corpus will in turn facilitate the spouse inference procedure of user i. $\psi_{Neigh}(C_i, X_e)$ encodes $\mathbf{I}(C_i \in S_e)$, $\mathbf{I}(C_i \notin S_e)$. Features we explore also include whether C_i's twitter ID appears in e's corpus.

4.3 Training

We separately trained three classifiers regarding the three attributes. All variables are observed during training; we therefore take a feature-based approach to learning structure prediction models inspired by structure compilation (Liang et al., 2008). In our setting, a subset of the features (those based on network information) are computed based on predictions that will need to be made at test time, but are observed during training. This simplified approach to learning avoids expensive inference; at test time, however, we still need to jointly predict the best attribute values for friends as is described in section 4.4.

4.4 Inference

Job and Education Our inference algorithm for Job/Education is performed on two settings, depending on whether neighbor information is observed (NEIGH-OBSERVED) or latent (NEIGH-LATENT). Real world applications, where network information can be partly retrieved from all types of social networks, can always falls in between.

Inference in the NEIGH-OBSERVED setting is trivial; for each entity $e \in G_i$, we simply predict it's candidate attribute values using Equ.6.

$$z_{i,e}^k = \underset{z'}{\text{argmax}}\, \Psi(z', X_i, F_i^k) \quad (6)$$

For NEIGH-LATENT setting, attributes for each node along the network are treated latent and user attribute prediction depends on attributes of his neighbors. The objective function for joint inference would be difficult to optimize exactly, and algorithms for doing so would be unlikely to scale to network of the size we consider. Instead, we use a sieve-based greedy search approach to inference (shown in Figure 3) inspired by recent work on coreference resolution (Raghunathan et al., 2010). Attributes are initialized using only text features, maximizing $\Psi_{text}(e, X_i)$, and ignoring network information. Then for each user we iteratively re-estimate their profile given both their text features and network features (computed based on the current predictions made for their friends) which provide additional evidence.

In this way, highly confident predictions will be made strictly from text in the first round, then the network can either support or contradict low confidence predictions as more decisions are made. This process continues until no changes are made at which point the algorithm terminates. We empirically found it to work well in practice. We expect that NEIGH-OBSERVED performs better than NEIGH-LATENT since the former benefits from gold network information.

Spouse For Spouse inference, if candidate entity e has no correspondent twitter account, we directly determine $z^k_{i,e} = \text{argmax}_{z'} \Psi(z', X_i)$ from text features. Otherwise, the inference of $z^k_{i,e}$ depends on the z^k_{e,C_i}. Similarly, we initialize $z^k_{i,e}$ and z^k_{e,C_i} by maximizing text factor, as we did for Education and Job. Then we iteratively update z^k given by the rest variables until convergence.

5 Experiments

In this Section, we present our experimental results in detail.

5.1 Preprocessing and Experiment Setup

Each tweet posting is tokenized using Twitter NLP tool introduced by Noah's Ark[14] with # and @ separated following tokens. We assume that attribute values should be either name entities or terms following @ and #. Name entities are extracted using Ritter et al.'s NER system (2011). Consecutive tokens with the same named entity tag are chunked (Mintz et al., 2009). Part-of-speech tags are assigned based on Owoputi et al's tweet POS system (Owoputi et al., 2013).

Data is divided in halves. The first is used as training data and the other as testing data.

[14] https://code.google.com/p/ark-tweet-nlp/downloads/list

	Education	Job
AFFINITY	74.3	14.5

Table 3: Affinity values for Education and Job.

5.2 Friends with Same Attribute

Our network intuition is that users are much more likely to be friends with other users who share attributes, when compared to users who have no attributes in common. In order to statistically show this, we report the value of AFFINITY defined by Mislove et al (2010), which is used to quantitatively evaluate the degree of HOMOPHILY in the network. AFFINITY is the ratio of the fraction of links between attribute (k)-sharing users (S_k), relative to what is expected if attributes are randomly assigned in the network (E_k).

$$S_k = \frac{\sum_i \sum_{j \in F_i^k} \mathbf{I}(P_i^k = P_j^k)}{\sum_i \sum_{j \in F_i^k} \mathbf{I}}$$
$$E_k = \frac{\sum_m T_m^k (T_m^k - 1)}{U^k (U^k - 1)} \quad (7)$$

where T_m^k denotes the number of users with m value for attribute k and $U^k = \sum_m T_m^k$. Table 3 shows the affinity value of the Education and Job. As we can see, the property of HOMOPHILY indeed exists among users in the social network with respect to Education and Job attribute, as significant affinity is observed. In particular, the affinity value for Education is 74.3, implying that users connected by a link in the network are 74.3 times more likely affiliated in the same school than as expected if education attributes are randomly assigned. It is interesting to note that Education exhibits a much stronger HOMOPHILY property than Job. Such affinity demonstrates that our approach that tries to take advantage of network information for attribute prediction of holds promise.

5.3 Evaluation and Discussion

We evaluate settings described in Section 4.2 i.e., GLOBAL setting, where user-level attribute is predicted directly from jointly feature space and LOCAL setting where user-level prediction is made based on tweet-level prediction along with different inference approaches described in Section 4.4,

i.e. NEIGH-OBSERVED and NEIGH-LATENT, regarding whether neighbor information is observed or latent.

Baselines We implement the following baselines for comparison and use identical processing techniques for each approach for fairness.

- **Only-Text:** A simplified version of our algorithm where network/neighbor influence is ignored. Classifier is trained and tested only based on text features.
- **NELL:** For Job and Education, candidate is selected as attribute value once it matches bag of words in the list of universities or companies borrowed from NELL. For Education, the list is extended by alias identification based on Freebase. For Job, we also fetch the name abbreviations[15]. NELL is only implemented for Education and Job attribute.

For each setting from each approach, we report the (P)recision, (R)ecall and (F)1-score. For LOCAL setting, we report the performance for both entity-level prediction (Entity) and posting-level prediction (Tweet). Results for Education, Job and Spouse from different approaches appear in Table 4, 5 and 6 respectively.

Local or Global For horizontal comparison, we observe that GLOBAL obtains a higher Precision score but a lower Recall than LOCAL(ENTITY). This can be explained by the fact that LOCAL(U) sets $z_{i,e}^k = 1$ once one posting $x \in L_i^e$ is identified as attribute related, while GLOBAL tend to be more meticulous by considering the conjunctive feature space from all postings.

Homophile effect In agreement with our expectation, NEIGH-OBSERVED performs better than NEIGH-LATENT since erroneous predictions in NEIGH-LATENT setting will have negative influence on further prediction during the greedy search process. Both NEIGH-OBSERVED and NEIGH-LATENT where network information is harnessed, perform better than **Only-Text**, which the prediction is made independently on user's text features. The improvement of NEIGH-OBSERVED over **Only-Text** is 22.7% and 6.4% regarding F-1 score for Education and Job respectively, which further illustrate the usefulness of making use of Homophile effect for attribute inference on online social media. It is also interesting to note the improvement much more significant in Education inference than Job inference. This is in accord with what we find in Section 5.2, where education network exhibits stronger HOMOPHILE property than Job network, enabling a significant benefit for education inference, but limited for job inference.

Spouse prediction also benefits from neighboring effect and the improvement is about 12% for LOCAL(ENTITY) setting. Unlike Education and Job prediction, for which in NEIGH-OBSERVED setting all neighboring variables are observed, network variables are hidden during spouse prediction. By considering network information, the model benefits from evident clues offered by tweet corpus of user e's spouse when making prediction for e, but also suffers when erroneous decision are made and then used for downstream predictions.

NELL Baseline Notably, NELL achieves highest Recall score for Education inference. It is also worth noting that most of education mentions that NELL fails to retrieve are those involve irregular spellings, such as HarvardUniv and Cornell_U, which means Recall score for NELL baseline would be even higher if these irregular spellings are recognized in a more sophisticated system. The reason for such high recall is that as our ground truths are obtained from Google plus, the users from which are mostly affiliated with decent schools found in NELL dictionary. However, the high recall from NELL is sacrificed at precision, as users can mention school entities in many of situations, such as paying a visit or reporting some relevant news. NELL will erroneously classify these cases as attribute mentions.

NELL does not work out for Job, with a fairly poor 0.0156 F1 score for LOCAL(ENTITY) and 0.163 for LOCAL(TWEET). Poor precision is expected for as users can mention firm entity in a great many of situations. The recall score for NELL in job inference is also quite low as job related entities exhibit a greater diversity of mentions, many of which are not covered by the NELL dictionary.

Vertical Comparison: Education, Job and Spouse Job prediction turned out to be much more difficult than Education, as shown in Tables 4 and 5. Explanations are as follows: (1) Job contains a much greater diversity of mentions than Education. Education inference can benefit a

[15] http://www.abbreviations.com/

		GLOBAL			LOCAL(ENTITY)			LOCAL(TWEET)		
		P	R	F	P	R	F	P	R	F
Our approach	NEIGH-OBSERVED	**0.804**	**0.515**	**0.628**	**0.524**	0.780	**0.627**	**0.889**	0.729	**0.801**
	NEIGH-LATENT	0.755	0.440	0.556	0.420	0.741	0.536	0.854	0.724	0.783
Only-Text	—-	0.735	0.393	0.512	0.345	0.725	0.467	0.809	0.724	0.764
NELL	—-	—-	—-	—-	0.170	**0.798**	0.280	0.616	**0.848**	0.713

Table 4: Results for Education Prediction

		GLOBAL			LOCAL(ENTITY)			LOCAL(TWEET)		
		P	R	F	P	R	F	P	R	F
Our approach	NEIGH-OBSERVED	**0.643**	**0.330**	**0.430**	**0.374**	**0.620**	**0.467**	**0.891**	**0.698**	**0.783**
	NEIGH-LATENT	0.617	0.320	0.421	0.226	0.544	0.319	0.804	0.572	0.668
Only-Text	—-	0.602	0.304	0.404	0.155	0.501	0.237	0.764	0.471	0.583
NELL	—-	—-	—-	—-	0.0079	0.509	0.0156	0.094	0.604	0.163

Table 5: Results for Job Prediction

		GLOBAL			LOCAL(ENTITY)			LOCAL(TWEET)		
		P	R	F	P	R	F	P	R	F
Our approach	—-	**0.870**	**0.560**	**0.681**	**0.593**	**0.857**	**0.701**	**0.904**	**0.782**	**0.839**
Only-Text	—-	0.852	0.448	0.587	0.521	0.781	0.625	0.890	0.729	0.801

Table 6: Results for Spouse Prediction

lot from the dictionary relevant feature which Job may not. (2) Education mentions are usually associated with clear evidence such as homework, exams, studies, cafeteria or books, while situations are much more complicated for job as vocabularies are usually specific for different types of jobs. (3) The boundary between a user working in and a fun for a specific operation is usually ambiguous. For example, a Google engineer may constantly update information about outcome products of Google, so does a big fun. If the aforementioned engineer barely tweets about working conditions or colleagues (which might still be ambiguous), his tweet collection, which contains many of mentions about outcomes of Google product, will be significantly similar to tweets published by a Google fun. Such nuisance can be partly solved by the consideration of network information, but not totally.

The relatively high F1 score for spouse prediction is largely caused by the great many of non-individual related entities in the dataset, the identification of which would be relatively simpler. A deeper look at the result shows that the classifier frequently makes wrong decisions for entities such as userID and name entities. Significant as some spouse relevant features are, such as love, husband, child, in most circumstances, spouse mentions are extremely hard to recognize. For example, in tweets "Check this out, @alancross, it's awesome `bit.ly/1bnjYHh`." or "Happy Birthday @alancross !". *alancross* can reasonably be any option among current user's friend, colleague, parents, child or spouse. Repeated mentions add no confidence. Although we can identify *alancross* as spouse attribute once it jointly appear with other strong spouse indicators, they are still many cases where they never co-appear. How to integrate more useful side information for spouse recognition constitutes our future work.

6 Conclusion and Future Work

In this paper, we propose a framework for user attribute inference on Twitter. We construct the publicly available dataset based on distant supervision and experiment our model on three useful user profile attributes, i.e., Education, Job and Spouse. Our model takes advantage of network information on social network. We will keep updating the dataset as more data is collected.

One direction of our future work involves exploring more general categories of user profile attributes, such as interested books, movies, hometown, religion and so on. Facebook would an ideal ground truth knowledge base. Another direction involves incorporating richer feature space for better inference performance, such as multi-media sources (i.e. pictures and video).

7 Acknowledgments

A special thanks is owned to Dr. Julian McAuley and Prof. Jure Leskovec from Stanford University for the Google+ circle/network crawler, without

which the network analysis would not have been conducted. This work was supported in part by DARPA under award FA8750-13-2-0005.

References

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

Edward Benson, Aria Haghighi, and Regina Barzilay. 2011. Event discovery in social media feeds. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 389–398. Association for Computational Linguistics.

Sergey Brin. 1999. Extracting patterns and relations from the world wide web. In *The World Wide Web and Databases*.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of twitter users in non-english contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Wash*, pages 18–21.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *ICWSM*.

Mark Craven, Johan Kumlien, et al. 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, pages 1–12.

Ido Guy, Naama Zwerdling, Inbal Ronen, David Carmel, and Erel Uziel. 2010. Social media recommendation based on people and tags. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201. ACM.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke S Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*, pages 541–550.

Jiwei Li and Claire Cardie. 2013. Timeline generation: Tracking individuals on twitter. *arXiv preprint arXiv:1309.7313*.

Percy Liang, Hal Daumé III, and Dan Klein. 2008. Structure compilation: trading structure for features. In *Proceedings of the 25th international conference on Machine learning*.

Wendy Liu and Derek Ruths. 2013. Whats in a name? using first names as features for gender inference in twitter. In *2013 AAAI Spring Symposium Series*.

Wendy Liu, Faiyaz Al Zamal, and Derek Ruths. 2012. Using social media to infer gender composition of commuter populations. In *Proceedings of the When the City Meets the Citizen Workshop, the International Conference on Weblogs and Social Media*.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel. 2010. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A machine learning approach to twitter user classification. In *ICWSM*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

Delip Rao and David Yarowsky. 2010. Detecting latent user properties in social media. In *Proc. of the NIPS MLSN Workshop*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Delip Rao, Michael J Paul, Clayton Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *ICWSM*.

K Ravikumar, Haibin Liu, J Cohn, Michael E Wall, Karin Verspoor, et al. 2012. Literature mining of protein-residue associations with graph rules learned

through distant supervision. *Journal of biomedical semantics*, 3(Suppl 3):S2.

Alan Ritter, Sam Clark, Mausam, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Alan Ritter, Luke Zettlemoyer, Mausam, and Oren Etzioni. 2013. Modeling missing data in distant supervision for information extraction.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465. Association for Computational Linguistics.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 721–729. Association for Computational Linguistics.

Fei Wu and Daniel S Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM.

Jaewon Yang and Jure Leskovec. 2013. Overlapping community detection at scale: A nonnegative matrix factorization approach. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 587–596. ACM.