Interactive Gender Inference in Social Media

Zhu Zhu[1,2] Jingjing Wang[1,2] Shoushan Li[1,2*] Guodong Zhou[1,2]

[1] Natural Language Processing Lab, School of Computer Science and Technology, Soochow University, China
[2] Collaborative Innovation Center of Novel Software Technology and Industrialization

{zhuzhu0020,djingwang,shoushan.li}@gmail.com
gdzhou@suda.edu.cn

Abstract. In this paper, we define a novel task named interactive gender inference, which aims to utilize interactive text to identify the genders of two interactive users. To address this task, we propose a two stage approach by well incorporating the dependency among the interactive samples sharing identical users. Specifically, we first apply a standard four-category classification algorithm to get a preliminary result, and then propose a global optimization algorithm to achieve better performance. Evaluation demonstrates the effectiveness of our proposed approach to interactive gender inference.

Keywords: gender Inference, social media

1 Introduction

The vigorous growth of social media in recent years, such as Twitter and Facebook, has produced an unprecedented amount of user-generated data. The open availability of such data provides an excellent opportunity to research on latent demographic features of online users. Gender inference is such a research issue which aims to identify the gender of a user. Due to its wide applications in social media analysis, gender inference has attracted more and more attention in recent years (Schler et al., 2006; Mukherjee et al., 2010; Tang et al., 2011; Ciot et al., 2013).

However, previous studies mainly focus on gender inference on a single user. In fact, social media not only provides a platform for a single user to publish his own experiences or opinions, but also offers an opportunity for the users in a community to interact, e.g. allowing a user to comment on a certain message. This interactive process normally yields some kind of interactive text which is helpful to infer the genders of the involved users. Let's consider following interactive text as an example:

E1: User **A**: *Thanks for your cool necklace for my birthday. Miss you.*
User **B**: *I miss you too, my dear wife. I'll be home soon.*

While it may be difficult to infer the gender of user **A** from the text by user **A** only, it is easy to infer the gender of user **A** from the text by user **B** due to the existence of "*my dear wife*", and also the gender of user **B**.

In this study, we focus on jointly inferring the genders of users involved in an interaction. For short, we name it interactive gender inference. Interactive gender

* Corresponding author

inference has additional applications, e.g. helping understand interpersonal communication mechanisms in sociology research (Dev et al., 2014) and driving AI research in human-machine interaction.

A straightforward approach to interactive gender inference is to cast it as a four-category classification problem, since four categories, i.e., *male to male*, *male to female*, *female to female* and *female to male* (denoted as *mm*, *mf*, *ff* and *fm*) can be naturally inferred from the interactive text. The problem with the four-category classification approach is that it ignores the dependency among the users and the interactive text. For instance, in one case, we may predict one sample from **A** to **B** as *mf*, while in another case, we may predict another sample from **A** to **C** as *ff*. Obviously, these two predictions are contradictory.

In this paper, we address interactive gender inference with the focus on all the comments from a user to another user as the interactive text to infer the genders of the involved two users. Specifically, motivated by the constraint that the gender of a user should keep the same in his all involved interactions, we propose a two-stage classification approach to interactive gender inference. In the first stage, we perform standard four-category classification to infer the genders of the involved two users. In the second stage, we propose a global optimization algorithm to benefit from the above constraint.

2 Related Work

In the last decade, most of related studies in gender inference deal with either blog text (Schler et al., 2006) or email text (Mohammad et al., 2011). For instance, Schler et al. (2006) exploit the difference in writing style and content between *male* and *female* bloggers to determine an author's gender. Similar to other classification tasks, most of such studies focus on exploring effective features to improve the performance (Mukherjee et al. 2010; Peersman et al. 2010; Gianfortoni et al. 2011).

More recently, with the rapid growth of social media, more and more researchers turn to micro-blog text. Burger et al. (2011) describe the construction of a large multilingual dataset labeled with the genders of Twitter users. Miller et al. (2012) identify the genders of Twitter users using Perceptron and Naive Bayes with selected *n*-gram features. Ciot et al. (2013) conduct the first assessment of the latent attribute inference in languages beyond English, focusing on gender inference of Twitter users.

3 A Two-Stage Approach

Generally, the users and user interactions in micro-blog can be represented as a graph, i.e., $G=(V,E)$ where V is the set of $|V|=N$ users and $E \subset V \times V$ is a set of $|E|=M$ interactive edges among users. Here, an interactive edge $e_{i,j} \in E$ is directed and represented by the comment text from $v_i \in V$ to $v_j \in V$. In this paper, our objective is to learn a model to infer the interactive gender of each edge $e_{i,j} \in E$, namely interactive gender.

In this paper, interactive gender is represented as a triple $(e_{i,j}, Rr_{ij}, Pp_{ij})$, where $e_{i,j} \in E$ is an edge; $Rr_{ij} \in YY$ is the label associated with the edge $e_{i,j}$ and

$YY = \{mm, mf, fm, ff\}$ is the category set. Pp_{ij} is the probability vector obtained by an algorithm for inferring the interactive gender. Generally, an interactive gender could also be inferred from the respective user gender, defined as a triple (v_i, r_i, p_i), where $v_i \in V$ is a node; $r_i \in Y$ is a label associated with the node, and $Y = \{male, female\}$. For convenience, we assume $r_i = 1$ if the label is *male*; Otherwise, $r_i = 0$. p_i is the probability vector.

3.1 Stage 1: Four-category Classification

Formally, the objective of four-category classification is illustrated as follows:

$$f(x) \to y \quad \text{Where } y \in YY \text{ and } YY = \{mm, mf, fm, ff\} \tag{1}$$

In this first stage, this predictor is used to determine the interactive genders of all the edges in the test data. That is to say, we obtain all the triples $(e_{i,j}, Rr_{ij}, Pp_{ij})$ where $e_{i,j}$ is a sample in the test data. Specifically, Pp_{ij} contains the probabilities of the sample belonging to each category, i.e.,

$$Pp_{ij} = <p(Rr_{ij} = mm), p(Rr_{ij} = mf), p(Rr_{ij} = ff), p(Rr_{ij} = fm)> \tag{2}$$

3.2 Stage 2: Global Label Optimization

The objective of global label optimization is to minimize the difference between the true gender label of each user and the inferred gender from the first stage. In general, each user node is involved in two types of edges: (a) the node appears in the left side of the edge; (b) the node appears in the right side of the edge. **Figure 1** shows the node sets connected to the node v_i.

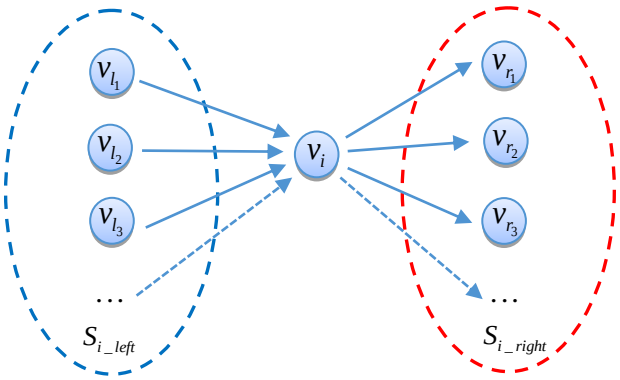

Figure 1. The node sets connected to the node v_i where S_{i_right} denotes the node set that contains the nodes from the right side and S_{i_left} denotes the node set that contains the nodes from the left side.

(a) When user v_i appears as the left node in the edge, the overall difference between the true label r_i and the inferred gender label is given as follows:

$$\sum_{j \in S_{i\_right}} \left(r_i - \tilde{r}_{ij} \right)^2 \tag{3}$$

Where $\tilde{r}_{ij}$ is inferred from the interactive gender of the edge $e_{i,j}$ according to following inference rule

$$\tilde{r}_{ij} = \begin{cases} 1 & \text{if } Rr_{ij} = mm \text{ or } Rr_{ij} = mf \\ 0 & \text{Otherwise} \end{cases} \tag{4}$$

(b) When user v_i appears as the right node in the edge, the overall difference between the true label r_i and the inferred gender label is given as follows:

$$\sum_{k \in S_{i\_left}} \left(r_i - \tilde{r}_{ki} \right)^2 \tag{5}$$

Where $\tilde{r}_{ki}$ is inferred by the interactive gender of the edge $e_{k,i}$ according to following inference rule,

$$\tilde{r}_{ki} = \begin{cases} 1 & \text{if } Rr_{ki} = mm \text{ or } Rr_{ki} = fm \\ 0 & \text{Otherwise} \end{cases} \tag{6}$$

When all users with both left and right connections are considered, our objective is to minimize the overall difference as follows,

$$\min \sum_{i=1}^{N} \left(\sum_{j \in S_{i\_right}} \left(r_i - \tilde{r}_{ij} \right)^2 + \sum_{k \in S_{i\_left}} \left(r_i - \tilde{r}_{ki} \right)^2 \right) \tag{7}$$

where *N* is the number of all users. To solve the above optimization problem, we first compute the partial derivative of r_i

$$\sum_{j \in S_{i\_right}} (2 \cdot r_i - 2 \cdot \tilde{r}_{ij}) + \sum_{k \in S_{i\_left}} (2 \cdot r_i - 2 \cdot \tilde{r}_{ki}) \tag{8}$$

Then, the best value of r_i to minimize formula (8) is the one making formula (8) equal 0. Thus, we get

$$r_i = \frac{1}{\left| S_{i\_right} \right|} \cdot \sum_{j \in S_{i\_right}} \tilde{r}_{ij} + \frac{1}{\left| S_{i\_left} \right|} \cdot \sum_{k \in S_{i\_left}} \tilde{r}_{ki} \tag{9}$$

From this formula, we can see that the optimization value of r_i is the average value of the gender labels inferred from both the left and right connected nodes.

The final decision on the user gender is made according to the value of r_i, i.e.,

$$\text{assign } v_i \to male \text{ if } r_i > 0.5 \text{ Otherwise } v_i \to female \tag{10}$$

Subsequently, the final decision on the interactive gender is made according to the user gender.

4 Experimentation

Data Setting

The data is collected from Sina Micro-blog (http://weibo.com/), one of the most famous Micro-blog platforms in China. From the website, we crawl user homepage which contains user messages (e.g. *name*, *gender*, *verified type*), and corresponding

comments. Overall, we get a data set of 53,675 users. From this data set, we select 20191 users as the training group and 9339 users as the test group (these two groups have no interactions between each other). Furthermore, we omit those users with less than 10 comments. Table 1 shows the statistics about the final data set.

#	Training Data	Test Data
mm	2883	1109
mf	4462	1599
ff	10954	3395
fm	4596	1591
Total	22895	7694

Table 1. Statistics about the data set

Features

Each interactive sample is treated as a bag-of-words and transformed into a binary vector encoding the presence or absence of one word feature. To get the word features in Chinese text, we use ICTCLAS (http://www.ictclas.org/ictclas_download.aspx) to perform word segmentation on the Chinese text. Apart from these basic features, we include two kinds of complex features, F-measure and POS sequence pattern features, which yield the state-of-the-art performance in gender inference (Mukherjee and Liu, 2010).

Classification Algorithm

We use the maximum entropy (ME) algorithm which is implemented with the public tool, Mallet Toolkits (http://mallet.cs.umass.edu/).

Evaluation Measurement

The performance is evaluated by both the F1-score in each category and the macro average of all F1-scores (Macro-F1).

Experimental Results

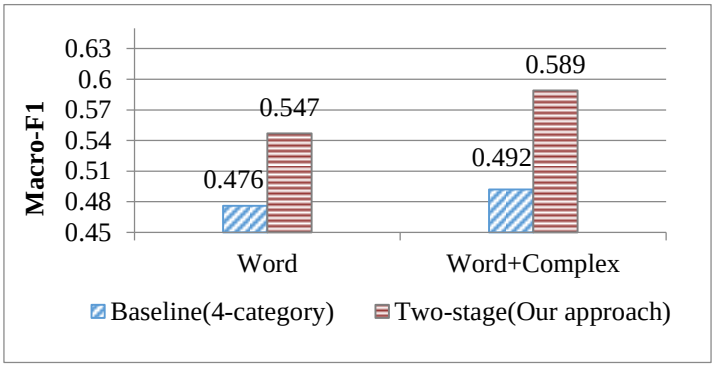

Figure 2: Performance comparison on interactive gender inference

Figure 2 compares the performance, where **Baseline** (4-category) means the four-category classification approach, i.e., the one employed in the first stage of our approach. From the figure, we can see that complex features, such as F-measure and POS sequence pattern features, significantly improve the performance. We can also see that our two-stage approach significantly outperforms the **Baseline** approach consistently. This indicates the dependency of interactive texts in user genders and the effectiveness of our global optimization algorithm.

5 Conclusion

In this paper, we address interactive gender inference with a two-stage approach. In the first stage, we utilize a standard four-category classification method to perform preliminary prediction. In the second stage, we propose a global optimization algorithm to benefit from the dependency among interactive texts in user genders. Evaluation on a large data set from Micro-blog platform shows the effectiveness of our two-stage approach over a strong baseline.

Acknowledgments

This research work has been partially supported by three NSFC grants, No. 61273320, No.61375073, No.61331011, and Collaborative Innovation Center of Novel Software Technology and Industrialization.

Reference

1. Burger J. and J. Henderson and G. Kim and G. Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of EMNLP-11*, pp. 1301–1309.
2. Ciot M., M. Sonderegger and D. Ruths. 2013. Gender Inference of Twitter Users in Non-English Contexts. In *Proceedings of EMNLP-13*, pp. 1136–1145.
3. Dev H., M. Ali, and T. Hashem. 2014. User Interaction Based Community Detection in Online Social Networks. In *Proceeding of DASFAA-14*, pp.296-310
4. Gianfortoni P., D. Adamson and C. Ros é. 2011. Modeling of Stylistic Variation in Social Media with Stretchy Patterns. In *Proceedings of EMNLP-11*, pp. 49–59.
5. Mukherjee A. and B. Liu. 2010. Improving Gender Classification of Blog Authors. In *Proceedings of EMNLP-10*, pp. 207-217.
6. Miller Z., B. Dickinson and W. Hu. 2012. Gender Prediction on Twitter Using Stream Algorithms with N-Gram Character Features. *International Journal of Intelligence Science*, Vol. 2, No. 4, pp.143-148.
7. Mohammad S. and T. Yang. 2011. Tracking Sentiment in Mail: How Genders Differ on Emotional Axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis(2011)*, pp.70-79.
8. Peersman C., W. Daelemans, L. Van Vaerenbergh. 2011. Predicting Age and Gender in Online Social Networks. In *Proceedings of SMUC-11*, pp. 37-44.
9. Schler J., M. Koppel, S. Argamon and J. Pennebaker. 2006. Effects of Age and Gender on Blogging. In *Proceedings of AAAI-06*, pp. 199-205.
10. Tang C., K. Ross, N. Saxena, and R. Chen. 2011. What's in a Name: A Study of Names, Gender Inference, and Gender Behavior in Facebook. In *Proceedings of DASFAA Workshops, 2011, LNCS 6637*, pp. 344-356.