I.J. Engineering and Manufacturing, **2011, 6, 14-24**
Published Online December 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2011.06.03

Available online at http://www.mecs-press.net/ijem

Automatically Extracting Name Alias of User from Email

Meijuan Yin, Xiao Li, Junyong Luo, Xiaonan Liu, Yongxing Tan

Information Science and Technology Institute, Zhengzhou 450002, China

Abstract

Mining user identity information from emails is an important research topic in email mining. Most approaches extract an email user's name only from the header of an email, but there are often many name information in the body of emails, which are usually more suitable for representing the sender's or recipient's identity. This paper focuses on the problem of extracting email users' name aliases in the body of plain-text emails. After locating and extracting salutation and signature blocks from email bodies, we can identify the potential aliases in the salutation and signature lines, which can be directly related with the email addresses in email headers, by using named entity recognition(NER) tools. To verify and amend the potential aliases that were identified by NER tools, we propose a novel approach to extract aliases in the salutation and signature lines based on name boundary word template built on the characteristics of alias neighboring words. Results on the public subset of the Enron corpus indicate that the approaches presented in this paper can efficiently extract user's aliases from email bodies.

Index Terms: Emails; Alias Extraction; Entity resolution; Salutation and signature blocks; Name boundary word template

1. Introduction

The popularity of the Internet makes people can communicate in many ways, such as Email, Blog, MSN and OICQ et al. To protect privacy, people usually use names different from their real-world names (denoted as aliases) when communicating with each other by Internet in a period of time. However, a user customarily employs a relatively fixed alias. So aliases can describe user identities to a certain extent. Mining user identities in network communicating is a popular research topic of data mining. This technique can be used in many network applications, such as identity recognition, information retrieval, social network analysis. This paper focuses only on email communication and studies the problem of extracting aliases of email users from email corpus.

Most approaches extract a user's name only from the header of an email, but there are often many names in the body of emails, which are usually fitter for representing the sender's or recipient's real identity. This paper focuses on the problem of extracting email users' names in the body of emails. Alias in the paper is a general

designation for formal name and informal name, such as anonym, nickname, short name and so on. In email bodies, only aliases that appear in the salutation and signature lines can be directly related with the email addresses in the header of emails. To effectively extract salutation and signature block from the body of an email, We propose the salutation and signature blocks locating algorithm based on statistical and rules restricted methods, which is presented in our former work [11]. Named entity recognition(NER) or part-of-speech tagging tools can be used to identify the potential aliases in the salutation and signature lines. But the identified aliases may be half-baked or there are still some potential aliases that can't be correctly identified. So we propose a novel approach to extract aliases in the salutation and signature lines. Using name boundary word template built on the characteristics of alias neighboring words, we verify and amend the potential aliases that were identified by NER tools. Results on the public subset of the Enron collection indicate that the approaches presented in this paper can efficiently extract user's aliases from the body of emails.

The remainder of this paper is organized as follows. Section 2 reviews earlier approaches to alias detection and email user identity modeling. The system framework to extract users' aliases from salutation and signature blocks in email bodies are introduced in section 3. The algorithm to extract aliases in salutation and signature blocks proposed in this paper is introduced in detail in section 4. In section 5, the method is evaluated on the public subset of the Enron collection. Results of our approach are concluded and discussed in section 6.

2. Related Work

Our approach of extracting aliases in emails relates to but different with the more general problem referred as "Entity Resolution." Entity resolution is generically defined as a process of determining the mapping from references (e.g., names, phrases) observed in data to real-world entities (e.g., persons, locations). Although much has been done on entity resolution [1], extracting aliases of personal names has not received enough attention. D. Bollegala et al. [2, 3, 4, 5] exploit trained models to extract candidate aliases of a given real personal name from Web pages. The approach can extract an entity's aliases, but the extracted aliases can't be associated with the email address of the entity. We restricted our work to the email collection. And as the email body is unstructured, it is difficult to elicit aliases from email bodies via model-based methods.

Identifying aliases of an email user is important for name reference resolution and entity's identity modeling in emails. C. Bird et al. [6] study the problem of correctly relating aliases and email addresses that belong to the same entity by clustering. They extract (alias, address) pairs from the header of emails and cluster them by the similarity between the pairs. C. Diehl et al. [7] firstly explore the problem of resolving personal name references in the full email including the message body. They build email communication social network based on the email sender-recipient relationship, and resolve the personal name references by using header-based traffic analysis techniques. Neither of those two studies reported the difficult problem of extracting users' name aliases from email bodies.

T. Elsayed et al. [8, 9, 10] also address the problem of resolving personal name references in the full email including the message body. They regard the email address as the key attribute to describe an entity identity, elicit names related to the email address from the header, the salutation and signature of an email, and resolve the personal name references by building the entity identity model. The method to partition the content and signature in the body of an email by using blank lines in Elsayed's research is very simple but effect for emails with normal bodies. When the body of an email is not consistent with the standard format, the method only using blank lines does not work well. Besides, to extract aliases from the salutation and signature, they do not directly use NER tools, but remove stop words and invalid sentences via a set of simple rules, compare the remaining lines with the user name in the email address, and select a whole line with some words similar to the user name as an alias of the user name.

Since above researches suffer from precision in extracting aliases of email users from the full email message, this paper is to propose a novel method to accurately and efficiently extract email user's aliases from email bodies.

3. Alias Extraction System

A. Outline

We use the email address to stand for an email user and all names related with the address are defined as the user's name aliases, which include his formal names and informal names such as anonym, nickname, short name and so on. The target of our alias extraction system is extracting all name aliases of each email user from email corpus, to provide alias information for alias authority analysis and identity identifying. The framework of email users' name alias extraction system is outlined in Fig. 1, composed of three modules: Alias Extraction in Email Header, Salutation and Signature Blocks Locating and Alias Extraction in Blocks. Using a set of email corpus as input, the system outputs all aliases of each email user appearing in the email corpus in the form of (email, alias) pairs. Firstly, we elicit (email, alias) pairs of the sender and recipient s from the address fields such as From, To, Cc and Bcc in email headers. Then locate and extract salutation and signature blocks from the body. In the end, we use NER tools to recognize candidate aliases in the extracted blocks, then related them to the email addresses elicited from the header and generate new (email, alias) pairs. The design of our alias extraction system is described in detail in the next section.

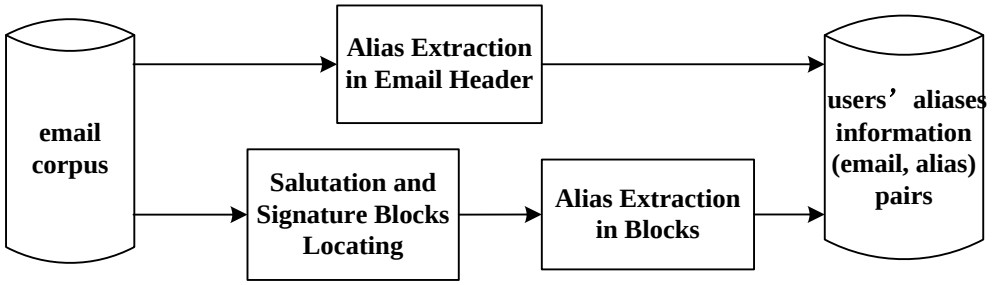

Fig. 1 Framework of email user's name alias extraction system

B. Design of Alias Extraction System

For a common email message, aliases that are found in two locations can be associated with the email address of a user. One is the "From, To, Cc and Bcc" headers where aliases appear with corresponding email addresses. The other is salutation and signature blocks in the email body, where aliases appear separately. A typical Enron email message is shown in Fig.2. Aliases extracted from "From, To, Cc and Bcc" headers can be directly associated with the corresponding addresses in headers. Aliases extracted from signature blocks should be associated with addresses in "From" headers, while aliases extracted from salutation blocks can only be associated with addresses in "To" headers. In the headers, aliases of email users appear in a fixed style, so we can easily elicit a (email, alias) pair by matching the strings of '<', '>' and ' " ' , ' " '. But, in the bodies, aliases associated with users' addresses appear not in a fixed style. And to extract them, we must resolve two key problems: one is the salutation and signature blocks locating and the other is aliases identifying and extracting accurately. So we propose Salutation and Signature Blocks Locating Algorithm based on statistical and rules restriction methods (abbreviated to SSBLA), which basically resolving the first problem. And to make the extracted aliases more precise, we present a novel approach called Name Boundary Word Template based Alias Extracting Algorithm(abbreviated to NBWT_AEA) to extract aliases in the salutation and signature blocks, which greatly settling the second problem.

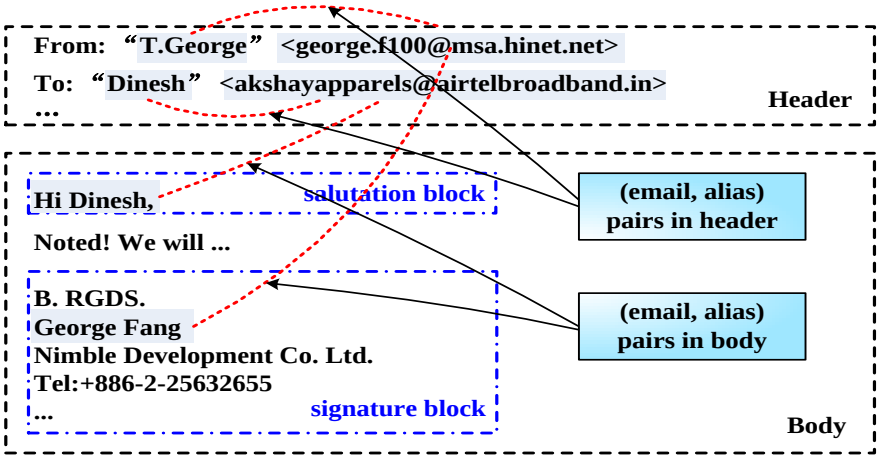

Fig. 2 Example of a typical email message

The input of our alias extraction system is email corpus and the output is users' alias information in the form of (email, alias) pairs. The primary steps of the system are illustrated in Fig.3. Firstly, directly extracting (email, alias) pairs from "From, To, Cc and Bcc" headers. Secondly, locating and eliciting salutation and signature blocks in the email body via SSBLA. The algorithm SSBLA is presented and evaluated in detail in our former work [11]. Thirdly, using NER or part-of-speech tagging tools to label texts in the blocks and obtain candidate aliases. Then, building and exploiting name boundary word template to verify and amend candidate aliases to valid aliases via BWTAEA, which is described in section 4. At last, associating each valid alias with the corresponding address in the "To" or "From" Header. In this way, we can finally extract all (email, alias) pairs and find all aliases of each user appearing in the email collection.

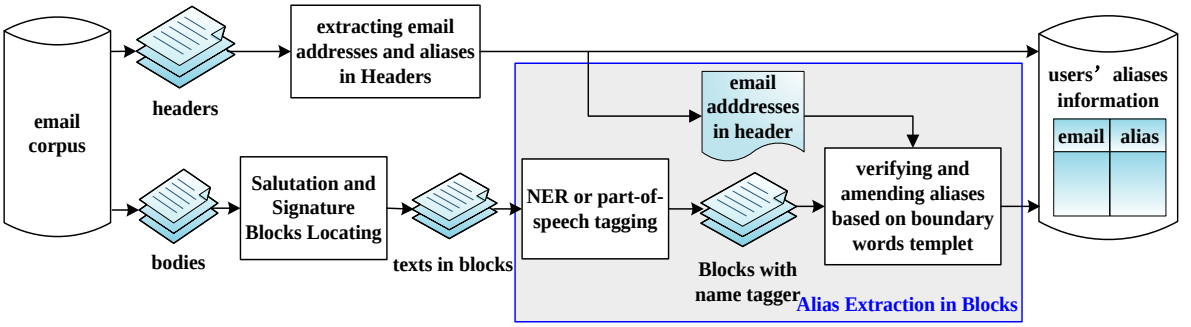

Fig. 3 Process flow of alias extraction system

4. Extracting Alias in salutation and signature blocks

In the part, we will describe the method to accurately extract aliases from the salutation and signature blocks located by SSBLA. We first introduce the definition of Name Boundary Word Template, and then present the Name Boundary Word Template based Alias Extracting Algorithm (NBWT_AEA) in detail.

1) Definition of Name Boundary Word Template

In the alias extraction system, after having located and elicited salutation and signature blocks from email bodies, we use part-of-speech tagging tools to label block texts and identify candidate aliases. There are some

relatively mature part-of-speech tagging tools in different languages. In this paper, taking emails in English or Chinese as examples, we present the method to build a common name boundary word template for aliases in salutation and signature blocks.

For English emails, we choose the well-known named entity recognition tool in English nature language process field, Named Entity Recognizer System Version1.1.1 [12] of Stanford University (abbreviated to NER). The label of names tagged by NER is a pair of labels "<PERSON>" and "</PERSON>", and between the pair of labels is a person name. For example, a result tagged by NER is "<PERSON>Jim Jarmusch</PERSON>", and the string "Jim Jarmusch" between the label pair "<PERSON>" and "</PERSON>" is an English person name. For Chinese emails, we select the famous part-of-speech tagging tool in Chinese nature language process field, ICTCLAS 3.0 of Chinese Academy of Sciences (CAS). The part-of-speech tagging label for a person name is "/nr", and it includes four sub-labels: "/nr1" is for a Chinese family name, "/nr2" is for a Chinese Christian name, "/nrj" is for a Japanese name, and "/nrf" is for a transliteration name. The Chinese characters before the label "/nr" is a potential person name.

By using named entity recognizing tools, we can identify most of the format names in salutation and signature blocks. However, the names appear in email bodies are usually informal names such as anonym, nickname, short name, honorific name and so on, which results in that by using named entity recognizing tools candidate aliases identified may not entire or even some aliases would not be identified. For example, the far-between Chinese name "Li Shuo" is labeled by ICTCLAS 3.0 as "Li/nr1 Shuo/ag ", based on which we can only extract the family name "Li", and for the Chinese nickname "Little Bean" labeled as "Little/a Bean/n", we can't extract any part of the nickname. According to above analysis, identifying aliases in email bodies only by named entity recognizing tools must induce inaccurate and missed aliases. In this paper, to improve the accuracy of extracting aliases, we build name boundary word template based on text features in email salutation and signature blocks, and then use the template to amend aliases identified by named entity recognizing tools.

The steps of building our name boundary word template are as follows.

a)Step 1: Defining the length of potential names.

Names in all kinds of languages commonly include two forms: one is a formal name, the other is an informal name, such as anonym, nickname, short name and so on. A formal name is usually composed of two parts (the family name or names and the Christian name), while an informal name may be composed of only one part (the family name or the Christian name) or two parts (the family name or the Christian name plus a salutation word to express respect, intimacy or title). So we define the length, the minimum length and the maximum length of a potential name according to the length of each part.

Definition 1: length, Min-length, Max-length of a potential name.

$l(x)$: is the length of word sequence x, that is the number of minimum language element in the sequence, e.g. the minimum language element in Chinese is Chinese character, in English is English word.

Several special instance of $l(x)$:

- when x is a punctuation, $l(x)=1$; $l(x)=0$ denote x doesn't exist;
- if n is a potential name, the length of name n is expressed as $l(n)$.

$L_{n\min}$: is the minimum possible length of name n;

$L_{n\max}$: is the maximum possible length of name n;

For example, in general, if n is a Chinese name, then $L_{n\max}=4$、$L_{n\min}=1$, and if n is an English name, then $L_{n\max}=3$、$L_{n\min}=1$.

b)Step 2: Building front and rear boundary words list according to the special context words neighboring a name in email salutation and signature blocks.

By analyzing a large amount of email messages in Enron email corpus [13] and referring to related information about greeting words in letters, we find many front and rear boundary words of names in email salutation and signature blocks, of which the most frequent words are shown in Table 1.

Definition 2: length, Min-length of a front or rear boundary word.

f : is the words sequence of a front boundary word, then the length of f is $l(f)$. And if f is a punctuation or special character, e.g. SPACE、CRLF, then $l(f)=1$;

r : is the words sequence of a rear boundary word, then the length of r is $l(r)$. And if r is a punctuation or special character, e.g. SPACE、CRLF, then $l(r)=1$;

$L_{f\max}$: is the maximum length of f, in English $L_{f\max}=2$;

$L_{r\max}$: is the maximum length of r, in English $L_{r\max}=1$.

Table 1 List of name boundary words in English email salutation and signature blocks

front boundary	Dear, My dear, Hi, "Hi,", Hello, "Hello,", Honorable, Hon. , Yours, Yours sincerely, Sincerely yours, Sincerely, Yours faithfully, Faithfully yours, Yours truly, Truly yours, Yours respectfully, Respectfully yours, cordially……
rear boundary	',', SPACE, CRLF……

c)Step 3: Defining name boundary word template in email salutation and signature blocks.

Definition 3: name boundary word template FNR1.

In email salutation and signature blocks, if there is a word sequence < fnr >, which satisfies that f is one of the front boundary words, r is one of the rear boundary words, n or part of n is a name labeled by NER tools, and $0<=l(f)<=L_{f\max}$, $0<=l(r)<=L_{r\max}$, $l(f)\times l(r)\neq 0$, $L_{n\min}<=l(n)<=L_{n\max}$, then named < fnr > as a name boundary word template in email salutation and signature blocks, and the temple is marked as FNR1.

The above template is only fit for names identified by NER tools. We must define another template for names that have not been identified by NER tools.

Definition 4: name boundary word template FNR2.

In email salutation and signature blocks, if there is a word sequence < fnr >, which satisfies that f is one of the front boundary words, r is one of the rear boundary words, n is an arbitrary word sequence, and $0<l(f)<=L_{f\max}$、$0<l(r)<=L_{r\max}$、$L_{n\min}<=l(n)<=L_{n\max}$, then named < fnr > as a name boundary word template in email salutation and signature blocks, and the temple is marked as FNR2.

2) Alias Extracting Algorithm

The basic idea of Name Boundary Word Template based Alias Extracting Algorithm is: if there is a name having been identified by NER tools in email salutation and signature blocks, then directly use name boundary word template FNR1 to amend the front and rear of the name, and get the corresponding alias to be extracted; otherwise, that is to say that there is no name having been identified by NER tools, employ name boundary word template FNR2 to locate the word sequence n whose front and rear boundary words can both be affirmed, and the word sequence n is the alias to be extracted.

Definition 5:

T : is the text of salutation or signature blocks in email bodies having been labeled by NER tools.

$w(i)$: is the i th minimum language element in T (e.g. the i th word in English text).

n : is the word sequence n labeled as a personal name by NER tools.

x : is the sequence number of n in text T , that is the sequence number of the first word of n in T .

The steps of the algorithm to amend a potential alias having been labeled as a name by NER tools based on name boundary word template FNR1 (abbreviated to FNR1A) are shown in Fig. 4.

The steps of the algorithm to extract a potential alias based on name boundary word template FNR2 (abbreviated to FNR2A) are shown in Fig. 5.

1. judge whether there is a rear boundary word after n in T ;
   if($\exists a,b$, satisfying (the word sequence $w(a)..w(a+b)$ is a rear boundary word in the boundary words list) && ($x+l(n) \le a < l(T)$) && ($0 \le b < L_{r\max}$) && ($a+b < l(T)$))
2. if true, then amend the rear of n ;
   { $temp$ = the word sequence $w(x)..w(a-1)$;
   if($l(temp) <= L_{n\max}$) $n = temp$; $l(n) = l(temp)$;
   if($l(temp) == L_{n\max}$) turn to step 5;
   }
3. judge whether there is a front boundary word before n in T ;
   if($\exists a,b$, (the word sequence $w(a)..w(a+b)$ is a front boundary word in the boundary words list) && ($0 \le a < x$) && ($0 \le b < L_{f\max}$) && ($a+b < x$)))
4. if true, then amend the front of n ;
   { $temp$ = the word sequence $w(a+b+1)..w(x+l(n)-1)$;
   if($l(temp) <= L_{n\max}$) $n = temp$;
   }
5. output n .

Fig. 4 Steps of algorithm FNR1A to extract alias based on template FNR1.

1. judge whether there is a front boundary word in T;
   if($\exists i, a$, satisfying (the word sequence $w(i)..w(i+a)$ is a front boundary word in the boundary words list) && ($0 \le i < l(T)$) && ($0 \le a < L_{f\max}$) && ($i + a < l(T)$))
2. if true, judge whether there is a rear boundary word in T;
   if($\exists j, b$, satisfying (the word sequence $w(j)..w(j+b)$ is a rear boundary word in the boundary words list) && ($i + a < j < l(T)$) && ($0 \le b < L_{r\max}$) && ($j + b < l(T)$))
3. if true, judge whether the front boundary word and the rear boundary word are neighbor in T;
   { if (j-(i+a)>1)
4. if the rear boundary word is not next to the front boundary word, then the word sequence between them is an aliases.
   { $temp$ = the word sequence $w(i+a+1)..w(j-1)$;
   if($l(temp) <= L_{n\max}$) $n = temp$;
   }
5. output n.

Fig. 5 Steps of algorithm FNR2A to extract alias based on template FNR2.

Name Boundary Word Template based Alias Extracting Algorithm (abbreviated to NBWT_AEA) is described in Fig. 6.

Algorithm: NBWT_AEA

Input: The text segment T of email salutation or signature blocks labeled by NER tools.
Output: the alias n in T
Begin:
1. initialize the potential alias word sequence n;
   $n = NULL$;
2. search for the word sequence n labeled as a name by NER tools in T, and then mark the word sequence number of n in T as x;
3. if($l(n) == 0$) turn to step 6;
4. if($l(n) == L_{n\max}$) turn to step 7;
5. call FNR1A(),amend the front and rear of n based on template FNR1;
   $n = FNR1A()$; turn to step 7;
6. call FNR2A(), extract the potential alias n that haven't been labeled as a name by NER tools based on template FNR2;
   $n = FNR2A()$;
7. output n.

End

Fig. 6 Steps of Name Boundary Word Template based Alias Extracting Algorithm.

The above algorithm NBWT_AEA can be used for emails in different languages. If only properly setting the list of name boundary words, the minimum language element, and the value of other related variables in above

definitions, then you can employ the algorithm to extract aliases from email salutation and signature blocks in other languages. In the experiment, we only test the algorithm on English emails, and the result is very good.

5. Evaluation

A)Datasets

In this section we analyze the methods described above. The experiments are carried out on the public Enron collection [13] published by Federal Energy Regulatory Commission(FERC) in 2003. It contains emails sent among 150 employees of the Enron corporation from October, 1998 to June, 2002. A part of those emails include salutation and signature blocks with different kinds of format, and the experiment results of our Salutation and Signature Blocks Locating Algorithm (SSBLA) [11] on those emails have shown a relatively high performance. The emails in the collection are stored in folders, each folder correspond to one user and include several sub-folders such as sent_mail folder, inbox folder, all_documents folder and so on. In the experiments we select the sent_mail folders of 20 users, which include 6065 emails, and from those folders we randomly choose 2000 emails which include names in "email-name" lists appended to the datasets. In those 2000 emails, after removing the quoted text from the email body, only 1672 emails have the text body, in which 358 emails include salutation blocks and 971 emails include signature blocks. About 3.2% of those emails with salutation or signature blocks do not include any names, and 1287 names appear in those salutation and signature blocks by labeling manually.

B)Experiments and Evaluation

We take the text segment of above 358 salutation blocks and 971 signature blocks labeled manually as the test dataset in our experiments. First, we extract all valid aliases from the dataset, associate them with the corresponding email addresses in "From" header or "To" header, and build the (email, name) pairs. Then, by comparing the pairs with the "email-name" lists and the results labeled manually, we can testify the validity of our Name Boundary Word Template based Alias Extracting Algorithm (NBWT_AEA).

We use two methods to extract aliases from salutation and signature blocks. In method 1, we directly employ Stanford NER tool to tag names in blocks and elicit names labeled by the tools as aliases associated with email users. In method 2, by using the algorithm NBWT_AEA we verify and amend the names labeled by method 1, and the results are taken as valid aliases associated with email users.

In the evaluation step, we think an alias of email user associated with his email address is correct if the alias string matches the labeled result. To evaluate the performance of the two alias extracting methods we use three measures: precision rate P, recall rate R and F1-measure F1, which are usually used to evaluate the performance in the Information Retrieval system. The formulas are defined as in (1):

$$P = n_{cname} / n_{ename} \;;\; R = n_{cname} / n_{rname} \;;\; F1 = 2PR/(P+R) \tag{1}$$

n_{ename} : is the number of aliases extracted from the email datasets by the alias extracting methods; n_{cname} : is the number of correct aliases in all of the extracted aliases; n_{rname} : is the total number of aliases labeled manually in the email datasets. Table 2 shows the evaluation results of two methods in above datasets.

Table 2 evaluation results of two alias extracting methods on above datasets

	n_{rname}	n_{ename}	n_{ename}	Precision(%)	Recall(%)	$F1$ measure(%)
method 1	1287	698	672	96.28	52.29	67.77
method 2	1287	1079	1053	97.59	81.82	89.01

Table 2 shows that our approach to extract aliases is much better than the method that only use NER tools to label aliases in both the precision and recall rate. Especially in recall rate, by analyzing the experiment result, we find that some inherent errors of Stanford NER system result in part of names can't be tagged correctly in above datasets, e.g. names such as Phillip and Theresa are labeled as place name, and names such as Darrell and Shelley are labeled as organization names. While as the appearance of most names in above datasets coincides with the name boundary word templates built in this paper, such as the form that a name is at the beginning of a line and next of the name is a comma, and that before the name is "Hi," which is at the beginning of a line and next of the name is CRLF. So the recall rate of our approach is much higher than that of method 1. But to names not matching the templates our method can't identify them still. Besides, by using the name boundary word templates parts of the half-baked names are amended correctly, which improves the alias extracting precision to a certain extent. However, there are some exceptional emails, such as names appear in signature blocks do not express contact information, and few signature blocks include more than one names. These exceptions have influence to the precisions of both two methods.

6. Conclusion

In this work we addressed the problem of automatically extracting aliases of email users from the full email message. The limitation of most existing related works is extracting aliases only from email address headers such as "From", "To" et al in email header, which makes the insufficient usage of email messages and the aliases extracted aren't overall. In allusion to this limitation, we proposed the novel approach to extract aliases of email sender and recipient from salutation and signature blocks in email bodies. After having located and extracted salutation and signature blocks from email bodies, we used Stanford NER system to identify the potential aliases in the salutation and signature lines, which can be directly related with the email addresses in email headers. To verify and amend the potential aliases that were identified by NER tools, we defined the name boundary word templates built on the characteristics of alias neighboring words to verify and amend the potential aliases identified by NER system, and thus obtained more valid and intact aliases. Results on the public subset of the Enron corpus indicate that the aliases extracting method presented in this paper can efficiently extract user's aliases from email bodies.

References

[1] Indrajit Bhattacharya and Lise Getoor. A latent dirichlet model for unsupervised entity resolution. In The SIAM International Conference on Data Mining (SIAM-SDM),Bethesda, MD, USA, 2006.

[2] D. Bollegala, Y. Matsuo, and M. Ishizuka. Disambiguating personal names on the web using automatically extracted key phrases. In Proc. of the 17th European Conference on Artificial Intelligence, pages 553-557, 2006.

[3] D. Bollegara, Y. Matsuo, and M. Ishizuka. Extracting key phrases to disambiguate personal names on the web. In Proc. CICLing 2006, 2006.

[4] D. Bollegala, T. Honma. Identification of Personal Name Aliases on the Web[A]. In: Proceedings of WWW 2008 Workshop on Social Web Search and Mining(SWSM 2008). Beijing, China, 2008.

[5] D. Bollegala, T. Honma, Y. Matsuo, and M. Ishizuka. Mining for personal name aliases on the web. In: Proceeding of the 17th international conference on World Wide Web, April 21-25, 2008, Beijing, China.

[6] Christian Bird, Alex Gourley and Anand Swaminathan. Mining Email Social Networks[A]. In: Proceedings of the 2006 international workshop on Mining software repositories [C]. Shanghai, China, 2006: 137-143.

[7] Chris Diehl, Lise Getoor, and Galileo Namata. Name reference resolution in organizational email archives. In Proceddings of SIAM International Conference on Data Mining,Bethesda, MD , USA, April 20-22 2006.

[8] T. Elsayed, Oard D W. Modeling Identity in Archival Collections of Email[A]. In: Proceedings of the Third Conference on Email and Anti-Spam[C]. Mountain View, California, USA, 2006.

[9] T. Elsayed, D. W. Oard, and G. Namata. Resolving personal names in email using context expansion. In Association for Computational Linguistics(ACL), 2008.

[10] T. Elsayed, G. Namata, L. Getoor, and D. W. Oard. Personal name resolution in email: A heuristic approach. Technical Report UMIACS LAMP-TR-150, University of Maryland, March 2008.

[11] M. Yin, J. Luo, D. Cao, X. Liu and M. Li. Automatically locating salutation and signature blocks in emails[A]. To be published in: Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'11) [C]. Shanghai, China, 2011.

[12] Stanford University. Named Entity Recognition System [EB/OL]. http://nlp.stanford.edu/software/stanford-ner-2009-01-16.tgz. 2009.

[13] The email collection of Enron Corporation [DB/OL]. http://www.cs.cmu.edu/~enron/. 2003.