Identification of Investigator Name Zones using SVM Classifiers and Heuristic Rules

Jongwoo Kim*, Daniel X. Le, George R. Thoma
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894, USA
*jongkim@mail.nih.gov

***Abstract*—The research reported in biomedical articles often involves large numbers of investigators at different institutions. To properly credit these investigators, an article's authors frequently name them together in some part of the article. These Investigator Names (IN) now constitute a required field in the MEDLINE® citation for the article. The automated extraction of these names is implemented in a system developed by a research group at the U.S. National Library of Medicine, consisting of three modules based on Support Vector Machine (SVM) classifiers and heuristic rules. The SVM classifiers label text blocks ("zones") that possibly contain Investigator Names, and the heuristic rules identify the actual zones. We collect eleven sets of word lists to train and test the classifiers, each set containing 100 to 56,000 words. Experimental results on online biomedical articles show a Precision of 0.90, 0.95 Recall, 0.92 F-Measure, and 0.99 Accuracy.**

Keywords— Investigator Names, MEDLINE, Support Vector Machine, heuristic rules, labeling, bibliographic information

I. INTRODUCTION

The U.S. National Library of Medicine (NLM) maintains MEDLINE, a bibliographic database that contains over 21 million citations for biomedical journal articles [1]. Each citation includes fifty-one fields of bibliographic data. NLM receives citations for many articles in XML format directly from journal publishers. However, key bibliographic information is often missing, requiring manual entry. The manual process to enter this data is time-consuming and error-prone. In addition, the volume of articles indexed for MEDLINE increases rapidly every year. We have therefore developed an automatic system called Publisher Data Review System (PDRS) to extract missing bibliographic information [2, 3, 4, and 5].

Investigator Names (IN) is one of the missing pieces of bibliographic information. These are names of investigators who collaborate in conducting research for the articles. They number above forty on average in an article, and sometimes over several hundred. There are three steps to extract IN in an article. First, divide an article into several zones (text blocks). Second, label zones containing IN. Third, extract IN from these labeled zones. In this paper, we will present an automatic labeling method for the second step.

Some algorithms commonly used for document labeling are Support Vector Machine (SVM) [6], the Naïve Bayes algorithm [7], and Conditional Random Fields (CRF) [8]. Naïve Bayes algorithm is used for spam emails [9] and Web document classification [10]. CRF is used for keyword extraction from documents [11] and segmenting/labeling documents [12]. SVM has been used to categorize newswire documents, Medical Subject Headings (MeSH) [13], Web documents [14], and the Reuters-21578 collection [15]. Since there are several variations (e.g., size, location, content) in the zones containing IN, supervised learning algorithms that handle nonlinear classification cases, are proper in this case. We, therefore, use SVM classifiers in this work. We proposed a prototype to label zones containing IN [16] and this paper presents the improved version of this prototype.

The remainder of this paper is organized as follows. Section II defines IN zones. The details of our method are presented in Section III. We discuss experimental results in Section IV, and show conclusions in Section V.

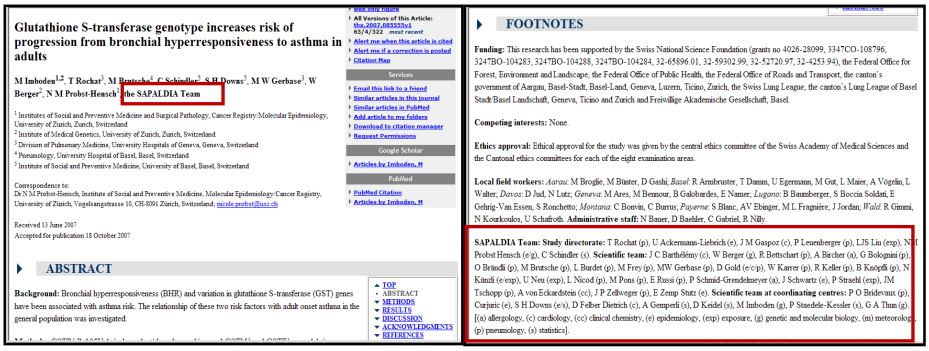
II. INVESTIGATOR NAME ZONE

In biomedical research, many investigators from different groups or organizations often collaborate in conducting the research. These groups or organizations also appear in the author or title zones, and the investigators affiliated with them would have their names listed somewhere else within the article, most likely toward the end, close to the references. We define the group/organization names as "Corporate Author". Investigators who are affiliated with a "Corporate Author" would have their names included in a zone (other than the author zone) in the article. We refer to names of the investigators as "Investigator Names (IN)" and zones containing these investigator names as "IN zones". Zones that do not contain IN are considered "Non-IN zones".

Fig. 1 shows an example of "Corporate Author" (CA) and its corresponding "IN zone." Fig. 1(a) shows the Corporate Author "SAPALDIA Team" located in the author zone and Fig. 1(b) shows the corresponding "IN zone" located in the footnote section. IN are usually grouped together in a single "IN zone" and appears in the end of articles located right above the reference section as shown in Fig. 1(b). However, there are exceptions as shown in Fig. 2. All "IN zones" in the figures are in red boxes. Fig. 2(a) shows an "IN zone" located next to an affiliation zone and Fig. 2(b) shows multiple "IN zones". Fig. 2(c) shows several "IN zones" that consist of only one investigator name per "IN zone". Fig. 2(d) shows a few (three) IN in an acknowledgment sentence with other information.

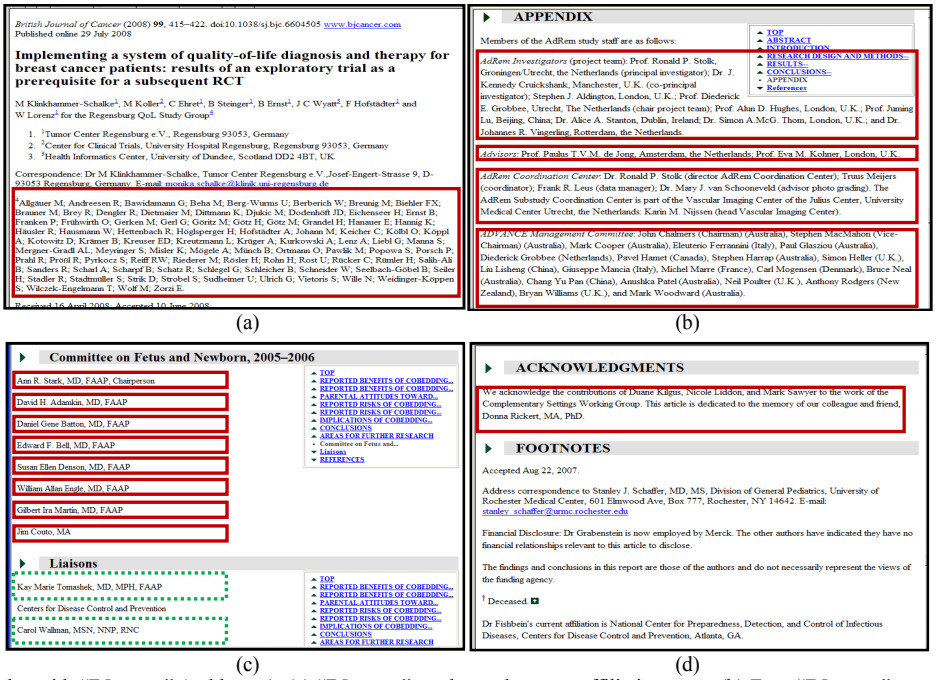
In addition, the contents of the "IN Zones" are expressed in different ways. Some "IN zones" contain only names (Fig. 2(a)), but others have names in addition to their affiliations (Fig. 2(b)) and other information (Fig. 2(d)). Some "IN zones"

contain few names (Figs. 2(c) and 2(d)) and some "IN zones" contain more other information than names (Fig. 2(d)). In Fig. 2(c), each "IN zone" consists of only one investigator name (red box) and its format is the same as the zones (green dashed box) in the Liaisons section. Therefore, information from neighboring zones, including section names, "Corporate Author" information, and zone locations within an article, are needed to identify "IN zones" correctly.



(a) (b)

Fig. 1: (a) An author zone with the Corporate Author "SAPALDIA Team." (b) The corresponding "IN zone" is located in Footnotes section.



(a) (b)

(c) (d)

Fig. 2. Examples of articles with "IN zones" (red boxes). (a) "IN zones" are located next to affiliation zone. (b) Four "IN zones" next to each other. (c). Eight "IN zones" next to each other followed by four "Non-IN zones". (d) Few (three) IN with other information in the "IN Zone."

III. PROPOSED APPROACH

A. *Features used for IN Zone*

"IN zones" are usually composed of two categories of words related to author names and/or affiliations. "Non-IN zones" are related to lexicons that are not related to the two word categories. Therefore, eleven word lists from MEDLINE are created as shown in Table 1. Based on the word lists, eighteen word features are collected for each zone as shown in Table 2. Among the features in the table, Journal name, Year, and Pagination are used to classify zones ("Non-IN zones") in Reference section that contain several author names.

Table I. WORD LISTS USED FOR LABELING AN "IN ZONE".

Word Lists	Explanation/Examples
Corporate Author words	Committee, Investigator, Group, etc.
Section name	Acknowledgment, Notes, etc.
Last name	Arison, Barret, Chay, Digman, Forbes, etc.
First name	Adam, Bent, Carol, Dave, Ema, Frank, etc.
Title	M.D., R.N., Ph.D., M.S., etc.
Affiliation	Department, School, Hospital, etc.
Grant Support	Grant, Support, Fund, etc.
Lexicon	Activation, case, delivered, project, etc.
Journal Name	Cell, Nature, JAMA, etc.
Unigram words from Journal Name	Abdominal, Biocell, Journal, ultrasonics, etc.
Bigram words from Journal Name	Abnormal Child, Circuits Systems, Journal for, The British, etc.

Table II. FEATURES USED FOR LABELING AN "IN ZONE".

Features	Explanation/Examples
Corporate Author name	Complete name
Common Corporate Author word	Committee, Investigator, Group, etc.
Section Name	Acknowledgment, Notes, etc. (excluding References)
References (Section Name)	The word "References" as title of that section.
Zone containing more than two words	
Journal Name	Existence of journal name
Year	Existence of year
Pagination	Existence of pagination
Section Name Zone	Check if the zone contains only section name
Frequency of Last and First Names	Proportion of these in a zone.
Frequency of All Names	Proportion of these in a zone.
Frequency of Common words in Title and Abstract zones	Proportion of these in a zone.
Frequency of Affiliation	Proportion of these in a zone.
Frequency of punctuation mark	Proportion of these in a zone.
Frequency of total words matched	Proportion of these in a zone.
Frequency of Grant Support	Proportion of these in a zone.
Normalized Frequency of All Names in a Zone	
Normalized Frequency of Number of Words in a Zone	

B. *Classifiers for IN Zones Used in the System*

The proposed system consists of two classifiers (Main-Classifier and Post-Classifier) and Heuristic Rules as shown in Fig. 3. The classifiers estimate candidate "IN zones" and heuristic rules merge split "IN zones" based on the results from previous classifier. We use SVM to train the classifiers.

The Main-Classifier is designed to classify a single "IN zone" containing several IN in an article using the features of the zone. Fig. 2(a) shows a similar "IN zone" for the classifier. The Post-Classifier is designed for classifying split "IN zones" (Fig. 2(b)) using the information from neighboring zones.

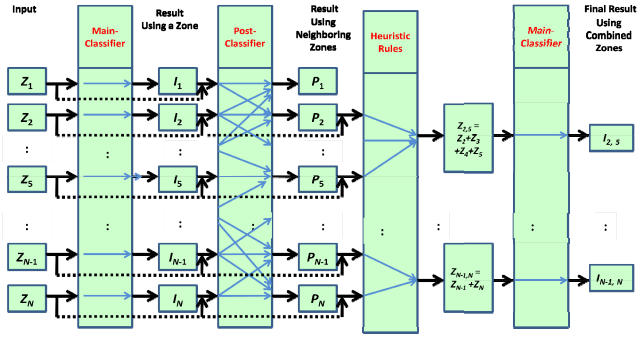


Fig. 3. The proposed "IN zone" classifier. I_i is the result of the Main-Classifier for zone i, P_i is the result of the Post-Classifier for zone i, $Z_{N-1,N}$ is a zone combined with Z_{N-1} and Z_N, and $I_{N-1,N}$ is the result of the Main-Classifier for the zone $Z_{N-1,N}$.

The main features of an "IN zone" are the names of investigators so the number of names identified in a zone plays an important role in recognizing a zone as an "IN zone." When an "IN zone" is divided into multiple zones, it is harder to label them as "IN zones". As a result, the Post-Classifier has to exploit information from neighboring zones to improve the classification rate in this case.

C. *Heuristic Rules used in the System*

It is hard to label all "IN zones" correctly using SVM classifiers especially for multiple "IN zones" in an article. Fig. 2(c) shows two candidates of "IN zone" groups. The first group zones (red boxes) belong to "IN zones" and the second group zones (green dashed boxes) do not belong to "IN zones" even if the two groups have the same characteristics. To distinguish the real "IN zones" group, we need to check the section name of the zone group. Therefore, the following rules are used to merge divided "IN zones."

Let $Z_{i,j}$ is a candidate "IN zones" group that starts from zone Z_i to Z_j.

If Z_{i-1} is the section name zone of "Acknowledgment", "Note", etc.,
  Label $Z_{i,j,}$ as "IN zones" and stop merging process for Z_{i-1}.
If Z_{i-1} has words in "Common Corporate Author word",
  Label $Z_{i,j}$ as "IN zones" and stop merging process for Z_{i-1}.
If Z_i has words in "Common Corporate Author word",
  Label $Z_{i,j,}$ as "IN zones" and stop merging process for Z_{i-1}.
If Z_{i+1} is the section name zone of "Acknowledgment", "Note", etc.,
  Label $Z_{i,j}$ as "IN zones" and stop merging process for Z_{i+1}.
If Z_{i-1} is the section name zone of "References", "Liaisons", etc.
  Label $Z_{i,j}$ as "Non-IN zones" and stop merging process for Z_{i-1}.

D. *Workflow of the System*

The following shows the workflow of the system (Fig. 3).

Let
I_i be the estimation result ($0 \le I_i \le 1$) of the Main-Classifier for zone i (Z_i),
P_i be the estimation result ($0 \le P_i \le 1$) of the Post-Classifier for zone i (Z_i) in an article, where $i = 1, 2,\ldots, N$
$I_{i,j}$ be the estimation result ($0 \le I_{i,j} \le 1$) of the Main-Classifier for $Z_{i,j}$ containing zones from Z_i to Z_j, where $0 \le i, j \le N$ and $i<j$.
W be the window size for merging "IN zones" using heuristic rules (W=2 in the case.).

First, estimate I_i of all Z_i in an article, where $i = 1, 2,\ldots, N$.

Second, sort I_i in descending order and select the first M zones that have the highest I_i values (Zh_i^m, $m = 1, 2,\ldots, M$ and $0 \le i \le N$).

Third, for each zone Zh_i^m, $m = 1, 2,\ldots, M$.
Step A. Set $h = i$, $k = i$, w=0, and threshold t (t =0.5 in this case).
Step B. Set $h = h - 1$.
  If ($P_h \ge t$)
    If (Z_h satisfy Heuristic Rules)
      Label Z_h as a "IN zone" and include Z_h in the Zh_i^m group.
      w=0.
    Else
      w++.
  Else
    w++.
  If ($W > w$)
    Go to Step B.
  Else
    Go to Step C.

Step C. Set $k = k + 1$.
If ($P_h \geq t$),
If (Z_k satisfy Heuristic Rules)
Label Z_k as a "IN zone" and include Z_k in the Zh_i^m group.
w=0.
Else
w++.
Else
w++.
If ($W > w$)
Go to Step C.
Else
Go to Step D.
Step D. Combine all "IN zones" from Z_h to Z_k to make a zone ($Z_{h,k}$) and estimate $I_{h,k}$ for $Z_{h,k}$.

Fourth, Sort $I_{h,k}$ in descending order where $0 \leq h,k \leq N$ and $h \leq k$.

Fifth, For each $Z_{h,k}$ from the highest $I_{h,k}$ value
If ($I_{h,k} > t$ and $Z_{h,k}$ satisfies the heuristic rules)
Select the $Z_{h,k}$ as "IN zones".
Stop.
Else
Continue.

The first step estimates the Main-Classifier results of all zones (N) in an article. The second step sorts the results (First) in descending order and picks the M zones that have the highest results. The third step, for each zone in the M zones, combines the split IN candidate zones using the Post-Classifier results and estimates the Main-Classifier results for the combined zone. The fourth step sorts the results (Third) for all M combined zones. The fifth step picks the combined zone with the highest results (Fourth) as IN zones.

IV. RESULTS AND DISCUSSION

Two sets of training data are collected for the Main-Classifier and Post-Classifier. For the training set of Main-Classifier, all "IN zones" in an article are combined into one zone. However, multiple "IN zones", as well as the combined zone, are used to train the Post-Classifier. Fig. 2(c) shows an example. The Main-Classifier uses the "IN zone" resulting from combining the eight separate "IN zones" for training. The Post-Classifier uses all eight individual "IN zones" as well as the combined "IN zone". For "Non-IN zones" for both classifiers, we randomly select zones from a set of training articles. To train the Post-Classifier, two neighboring zones (one zone before and one after an "IN zone") are used in this experiment.

For the two different classifiers used in the proposed system, we choose SVM classifiers and radial basis functions (RBF) as their kernel functions, implement them using the LIBSVM [17, 18] library, and used a sigmoid function to map the classifiers' results between 0 and 1.

To evaluate the performance of the proposed system, we use four different measures; Precision, Recall, F-Measure, and Accuracy.

There are not many articles that have IN. Nevertheless, we managed to collect 159 "IN zones" and 433 "Non-IN zones" to train the Main-Classifier and 448 "IN zones" and 530 "Non-IN zones" to train the Post-Classifier from 159 articles. To test the proposed system, we collect 216 different journal articles containing 323 "IN zones" and 46,217 "Non-IN zones".

Table I shows the test results. Among the 323 "IN zones", 308 "IN zones" are correctly labeled and 15 are under-labeled (false-negative errors). 36 "Non-IN zones" are over-labeled (false-positive errors) and 45,858 "Non-IN zones" are labeled correctly. Table II shows the performance based on the four different measures. Precision rate shows 0.90, Recall rate 0.95, F-Measure rate 0.92, and Accuracy rate 0.99. Table III shows results at the article level. Among 216 articles, 187 articles have all "IN zones" labeled correctly (86.57% of articles are labeled correctly). 20 articles (9.26%) are over-labeled and one article is under-labeled (0.46%) in the merging process using the heuristic rules. Eight articles are both under- and over-labeled (3.70%) because the Post-Classifier could not estimate one of the "IN zones" as having the highest confidence among the candidate "IN zones". Some author zones and reference zones are mislabeled as "IN zones" since they contain several names.

The proposed system produces some false positive and false negative errors. False-negative errors generated by the Main-Classifier cause false-positive errors in final results. In the case of false-positive errors, some are made by the Post-Classifier and by the heuristic rules. Fig. 4(a) shows a false-positive error. After the "IN zone" (red box) is labeled by the system, a "Non-IN zone" (green dashed box) was also labeled as "IN zone" since the zone contains several names. Fig. 4(b) shows a false-negative error. The two "IN zones" (red boxes) are labeled since there are several IN in the zones. However, the first zone (green dashed box) is not labeled as "IN zone" since there are few IN and all IN are foreign names. The names are not listed in the word lists for First and Last names.

In Fig. 4(a), the false-positive error can be resolved if the system uses the first word "correspondence" in the zone. In Fig. 4(b), the false-negative error can be resolved if the rules check the first words "Study Committee Co-chairs" since "Committee" belongs to "Common Corporate Author word" and the word is also used in "IN zones" in the article.

TABLE I. PERFORMANCE OF THE PROPOSED SYSTEM (216 ARTICLES USED)

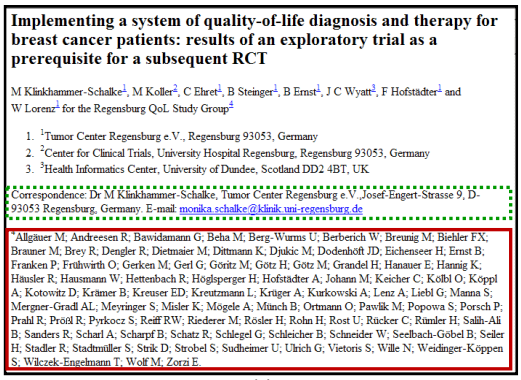
		True	False
IN Zone	(323)	308	15
Non-IN Zone	(45,894)	36	45,858

TABLE II. PERFORMANCE MEASURES FOR THE PROPOSED SYSTEM

	Proposed System
Precision	0.90
Recall	0.95
F-Measure	0.92
Accuracy	0.99

TABLE III. PERFORMANCE OF THE PROPOSED SYSTEM AT THE ARTICLE LEVEL

	Number of articles	Percent (%)
Total	216	
All "IN zones" labeled correctly	187	86.57
All "IN zones" labeled correctly and some "Non-IN zones" are over labeled	20	9.26
Some "IN zones" are under labeled	1	0.46
All "IN zones" are under labeled and some "Non-IN zones" are over labeled	8	3.70


Implementing a system of quality-of-life diagnosis and therapy for breast cancer patients: results of an exploratory trial as a prerequisite for a subsequent RCT

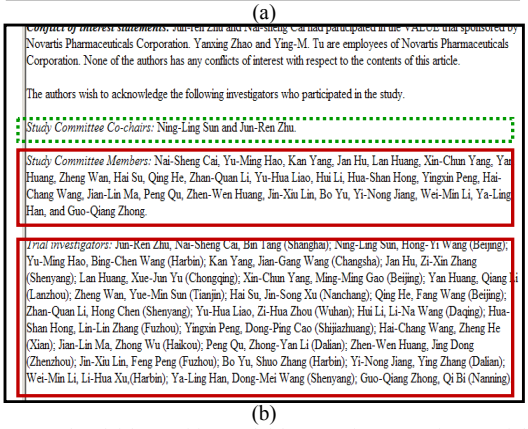
M Klinkhammer-Schalke[1], M Koller[2], C Ehret[1], B Steinger[1], B Ernst[1], J C Wyatt[3], F Hofstädter[1] and W Lorenz[1] for the Regensburg QoL Study Group[4]

1. [1]Tumor Center Regensburg e.V., Regensburg 93053, Germany
2. [2]Center for Clinical Trials, University Hospital Regensburg, Regensburg 93053, Germany
3. [3]Health Informatics Center, University of Dundee, Scotland DD2 4BT, UK

Correspondence: Dr M Klinkhammer-Schalke, Tumor Center Regensburg e.V., Josef-Engert-Strasse 9, D-93053 Regensburg, Germany. E-mail: monika.schalke@klinik.uni-regensburg.de

[4]Allgäuer M; Andreesen R; Bawidamann G; Beha M; Berg-Wurms U; Berberich W; Breunig M; Biehler FX; Brauner M; Brey R; Dengler R; Dietmaier M; Dittmann K; Djukic M; Dodenhöft JD; Eichenseer H; Ernst B; Franken P; Frühwirth O; Gerken M; Gerl G; Göritz M; Götz H; Götz M; Grandel H; Hanauer E; Hannig K; Häusler R; Hausmann W; Hettenbach R; Höglsperger H; Hofstädter A; Johann M; Keicher C; Kolbl O; Köppl A; Kotowitz D; Krämer B; Kreuser ED; Kreutzmann L; Krüger A; Kurkowski A; Lenz A; Liebl G; Manna S; Mergner-Gradl AL; Meyringer S; Misler K; Mögele A; Münch B; Ortmann O; Pawlik M; Popowa S; Porsch P; Prahl R; Pröll R; Pyrkocz S; Reiff RW; Riederer M; Rösler H; Rohn H; Rost U; Rücker C; Rümler H; Salih-Ali B; Sanders R; Scharl A; Scharpf B; Schatz R; Schlegel G; Schleicher B; Schneider W; Seelbach-Göbel B; Seiler H; Stadler R; Stadtmüller S; Strik D; Strobel S; Sudheimer U; Ulrich G; Vietoris S; Wille N; Weidinger-Köppen S; Wilczek-Engelmann T; Wolf M; Zorzi E.

(a)


Conflict of interest statements: Jun-Ren Zhu and Nai-Sheng Cai had participated in the VALUE trial sponsored by Novartis Pharmaceuticals Corporation. Yanxing Zhao and Ying-M. Tu are employees of Novartis Pharmaceuticals Corporation. None of the authors has any conflicts of interest with respect to the contents of this article.

The authors wish to acknowledge the following investigators who participated in the study.

Study Committee Co-chairs: Ning-Ling Sun and Jun-Ren Zhu.

Study Committee Members: Nai-Sheng Cai, Yu-Ming Hao, Kan Yang, Jan Hu, Lan Huang, Xin-Chun Yang, Yan Huang, Zheng Wan, Hai Su, Qing He, Zhan-Quan Li, Yu-Hua Liao, Hui Li, Hua-Shan Hong, Yingxin Peng, Hai-Chang Wang, Jian-Lin Ma, Peng Qu, Zhen-Wen Huang, Jin-Xiu Lin, Bo Yu, Yi-Nong Jiang, Wei-Min Li, Ya-Ling Han, and Guo-Qiang Zhong.

Trial investigators: Jun-Ren Zhu, Nai-Sheng Cai, Bin Tang (Shanghai); Ning-Ling Sun, Hong-Yi Wang (Beijing); Yu-Ming Hao, Bing-Chen Wang (Harbin); Kan Yang, Jian-Gang Wang (Changsha); Jan Hu, Zi-Xin Zhang (Shenyang); Lan Huang, Xue-Jun Yu (Chongqing); Xin-Chun Yang, Ming-Ming Gao (Beijing); Yan Huang, Qiang Li (Lanzhou); Zheng Wan, Yue-Min Sun (Tianjin); Hai Su, Jin-Song Xu (Nanchang); Qing He, Fang Wang (Beijing); Zhan-Quan Li, Hong Chen (Shenyang); Yu-Hua Liao, Zi-Hua Zhou (Wuhan); Hui Li, Li-Na Wang (Daqing); Hua-Shan Hong, Lin-Lin Zhang (Fuzhou); Yingxin Peng, Dong-Ping Cao (Shijiazhuang); Hai-Chang Wang, Zheng He (Xian); Jian-Lin Ma, Zhong Wu (Haikou); Peng Qu, Zhong-Yan Li (Dalian); Zhen-Wen Huang, Jing Dong (Zhenzhou); Jin-Xiu Lin, Feng Peng (Fuzhou); Bo Yu, Shuo Zhang (Harbin); Yi-Nong Jiang, Ying Zhang (Dalian); Wei-Min Li, Li-Hua Xu,(Harbin); Ya-Ling Han, Dong-Mei Wang (Shenyang); Guo-Qiang Zhong, Qi Bi (Nanning)

(b)

Fig. 4. An example of false-positive and false-negative errors in an article. (a) An "IN zone" (green dashed box) is over labeled as a "IN zone." (b) An "IN zone" (green dashed box) is under labeled as a "Non-IN zone."

V. CONCLUSIONS

This paper proposes an automated system consisting of two different SVM classifiers and heuristic rules to automatically label zones that contain Investigator Names in an online biomedical article.

We collect eighteen word features to train and test the two SVM classifiers. The Main-Classifier is used to label a zone with highest confidence as an "IN zone" and the Post-Classifier is used to label multiple "IN zones" using neighboring zones' information. The heuristic rules are also used to merge multiple "IN zones" to improve performance of the proposed system.

The proposed system shows relatively good performance. Precision is somewhat low (0.90), but, Recall and F-Measure are relatively high (0.96 and 0.92). The accuracy is also high (0.99). The Main-Classifier creates false-negative errors in articles with multiple "IN zones" and some false-positive errors result from author zones and references in the article. The Post-Classifier generates some false-positive errors in multiple "IN zones".

As a future task, we plan to use more features to eliminate errors made by the two classifiers, and add more robust rules to improve accuracy of merging process for the multiple "IN zones" cases.

ACKNOWLEDGMENT

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine, and Lister Hill National Center for Biomedical Communications.

REFERENCES

[1] http://www.nlm.nih.gov/pubs/factsheets/medline.html.

[2] J. Kim, D. X. Le, and G. R. Thoma, "Naïve Bayes and SVM Classifiers for Classifying Databank Accession Number Sentences from Online Biomedical Articles," *IS&T/SPIE's 22nd Annual Symposium on Electronic Imaging, San Jose, CA, 7534: 75340U-1-8 (2010).*

[3] J. Kim, D. X. Le, and G. R. Thoma, **"**Inferring Grant Support Types From Online Biomedical Articles," *Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems, Albuquerque, New Mexico. (2009).*

[4] J. Kim, D. X. Le, and G. R. Thoma, **"**Naïve Bayes Classifier For Extracting Bibliographic Information From Biomedical Online Articles," *Proc. International Conference on Data Mining, Las Vegas, Nevada, USA, II: 373-8 (2008).*

[5] "Technical Memorandum 484: Investigator Names," National Institutes of Health, National Library of Medicine, (2008).

[6] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, 121-167 (1998).

[7] D. D. Lewis, "Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval," *ECML*, The Tenth European Conference on Machine Learning, 4-15 (1998).

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", Proceedings of the 18th International Conference on Machine Learning (ICML01), Williamstown, MA, 282-289 (2001).

[9] D. Madigan, "Statistics and the war on spam," Statistics: A Guide to the Unknown, 4th Ed. (R. Peck, G. Casella, G. Cobb, R. Hoerl, D. Nolan, R. Starbuck and H. Stern, eds.), Thomson Brooks/Cole, Belmont, CA, 135–147 (2005).

[10] A. McCallum, and K. Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification," Proc. the AAAI-98 Workshop on Learning for Text Categorization, 577 (1998).

[11] C. Zhang, H. Wang, D. Wu, Y. Liao, and B. Wang, "Automatic Keyword Extraction from Documents Using Conditional Random Fields", Journal of Computational Information System, 1169-1180 (2008).

[12] S. Shetty, H. Srinivasan, and S. Srihari, "Segmentation and Labeling of Documents using Conditional Random Fields", Proceeding of Document Recognition and Retrieval IV, SPIE, San Jose, CA, 6500U-1-11 (2007).

[13] T. Joschims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proc. ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE, 137-142 (1998).

[14] E. Gabrilovich and S. Markovitch, "Text Categorization with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5", ICML'04, 321- 328 (2004).

[15] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," Proc. CIKM-98, 7th ACM International Conference on Information and Knowledge Management, Washington, 148–155 (1998).

[16] J. Kim, D. X. Le, and G. R. Thoma, "Combining SVM Classifiers to Identify Investigator Name Zones in Biomedical Articles", Proc. of SPIE, Document Recognition and Retrieval XIX, Vol. 8297, 829704, (2012).

[17] C. C. Chang, and C. J. Lin, "LIBSVM: a library for support vector machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, (2001).

[18] M. Johnson, "SVM.NET", Software available at http://www.matthewajohnson.org/index.html, (2008).