A stacked sequential learning method for investigator name recognition from Web-based medical articles

Xiaoli Zhang*, Jie Zou, Daniel X. Le, George Thoma
Communications Engineering Branch
National Library of Medicine, NIH, Bethesda MD 20814, USA
{zhangxiaol, jzou, danle, gthoma}@mail.nih.gov

ABSTRACT

"Investigator Names" is a newly required field in MEDLINE citations. It consists of personal names listed as members of corporate organizations in an article. Extracting investigator names automatically is necessary because of the increasing volume of articles reporting collaborative biomedical research in which a large number of investigators participate. In this paper, we present an SVM-based *stacked sequential learning* method in a novel application – recognizing named entities such as the first and last names of investigators from online medical journal articles. Stacked sequential learning is a meta-learning algorithm which can boost any base learner. It exploits contextual information by adding the predicted labels of the surrounding tokens as features. We apply this method to tag words in text paragraphs containing investigator names, and demonstrate that stacked sequential learning improves the performance of a non-sequential base learner such as an SVM classifier.

Keywords: Stacked Sequential Learning, Support Vector Machine (SVM), Named Entity Recognition

1. INTRODUCTION

Named entity recognition (NER) aims to identify names in categories such as person, organization, location, etc. in free text[1]. Since determining whether or not a particular word is a name, and identifying its entity type, also depends on the context of the word as well as the entity type of its neighbors, NER is often posed as a sequence classification problem and solved by sequential learning methods such as hidden Markov models (HMM), maximum entropy Markov models (MEMM), and conditional random fields (CRF). Bikel et al. used HMMs to represent a sentence (a sequence of words) in text and to find its most likely sequence of name-classes by using Viterbi algorithm[2]. McCallum et al. introduce MEMM, which combines the advantages of HMM and maximum-entropy models and performs better than either model on the task of FAQs segmentation[3]. Conditional random fields proposed by Lafferty et al. avoid the limitation of the label bias problem that exists in MEMM and other discriminative Markov models[4]. McCallum et al. applied CRFs to NER in news articles proposing a feature induction scheme to increase the conditional probability of a correct label sequence and improve the efficiency of the CRF[5]. The comparison of several sequence learners on the signature-detection problem with different features by Carvalho et al., and the evaluation of different sequence labeling algorithms applied to Part of Speech (POS) tagging and Optical Character Recognition tasks by Nguyen et al. indicate that the best sequential learner is very feature and application dependent[6, 7].

Containing 17 million citations to the medical journal literature, MEDLINE®, the flagship database of the U.S. National Library of Medicine, is a critical source of information for biomedical research and clinical medicine. With the rapid growth of the journal literature, automatic retrieval of bibliographic data such as article title, author name, affiliation, abstract, etc. is essential to replace the labor-intensive manual entry of MEDLINE citations. Beginning with journal issues published in 2008, personal names of those who are not entered as authors but belong to members of corporate organizations are required to be included in a new "Investigator Names" field in MEDLINE. The addition of these investigator names to MEDLINE is necessary because of the increasingly collaborative nature of biomedical research. Since not all of the investigators who collaborate are involved in preparing the manuscript or are included as authors, the inclusion of "Investigator Names" will allow retrieval of studies in which an individual took part.

*Xiaoli Zhang: zhangxiaol@mail.nih.gov, phone 1 -301- 435-3245.

In this paper, we propose an SVM-based stacked sequential learning method to automatically identify investigator names. The focus is on named entity recognition – tagging the first and last names of investigators from text paragraphs in medical articles. As noted previously, named entity recognition has been extensively applied to newswire, scientific articles, Web pages, and informal text like email[2, 5, 8, 9]. Extracting investigator names from medical articles belongs to a new domain of NER applications due to the specific appearance of investigator names in a text paragraph. Figure 1 shows four examples of text paragraphs containing investigator names captured from online HTML medical articles. Several names are marked with red boxes in each text paragraph. Investigator names appear in a group of word segments and can freely occur with degrees, affiliations, or locations. In contrast to paragraphs in general, the paragraphs that contain investigator names have the following properties: (1) no conventional grammatical structure exists as in a sentence; (2) there are rich variations of co-occurrence of named entities. Furthermore, due to the format of investigator names required in MEDLINE, the need to identify the first and last names of each investigator separately makes this task more challenging. By augmenting features using the predicted labels of nearby words, our stacked sequential learning method exploits the potential contextual information between named entities, and improves the performance of an SVM classifier for word tagging.

The remaining paper is organized as follows: In Section 2, we describe the features extracted for words in a text paragraph containing investigator names. Section 3 introduces the stacked sequential learning method. The experimental results are analyzed and the stacked sequential learning method is compared with the non-sequential learner SVM in Section 4. We summarize our work in Section 5.

2. FEATURE EXTRACTION

2.1 Preprocessing

The HTML article is first segmented into zones using geometric and text features. HTML zoning, discussed in our previous work, is a very useful preprocessing step for several information retrieval tasks[10, 11]. Next, text zones containing investigator names are extracted, which provide text strings as the input patterns to our investigator name recognition system.

In an NER task, a named entity can be a single word or a sequence of words. Investigator name recognition includes two steps: assigning to each word one of three categories - First Name, Last Name or neither ("Other"); uniting the labeled first and last name into a complete investigator name. The second step is based on heuristic rules which will not be discussed in this paper. We focus on the first step of tagging each one-token named entity which is a single word. As the final preprocessing step, the extracted investigator name text paragraphs are further segmented into words by either spaces or punctuations.

2.2 Feature description

Features used in our task include basic word-level features, dictionary features, and contextual features listed in Figure 2. All the features are in binary form with value 1 if a word has a particular feature, or else 0.

Basic features

The basic features are word-level features describing word case, punctuation, digit, and special characters. Digit patterns and digits mixed with characters are rarely names. A capital character (A-Z) followed by a period is very likely to be an initial which suggests a first name. A punctuation such as a hyphen may connect two first or last names.

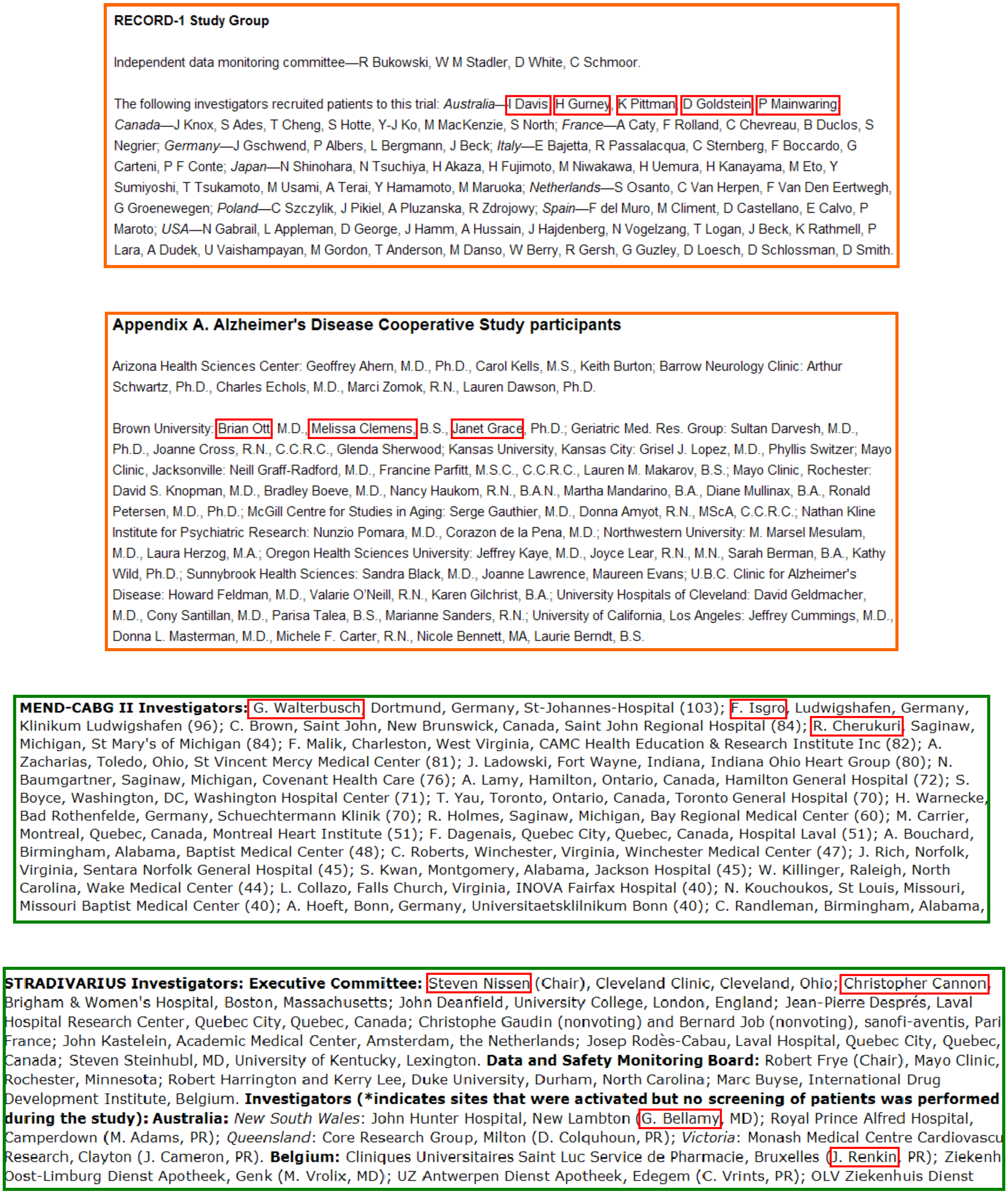
Dictionary features

We built First Name List, Last Name List, Affiliation Key Word List, Country Name List, US Canada State List and Title List from about 8 million medical articles in MEDLINE. If a candidate word matches an element in one of these lists, we set its corresponding dictionary feature to 1 indicating the possible entity type.

Contextual features

For each word, the features from neighboring words are also extracted to take advantage of the contextual dependencies between words.

A total of 16 basic and dictionary features are extracted and each word is represented by a binary feature vector. The SVM model is then learned from the binary feature vectors of the training words by adding different orders of contextual



RECORD-1 Study Group

Independent data monitoring committee—R Bukowski, W M Stadler, D White, C Schmoor.

The following investigators recruited patients to this trial: *Australia*—I Davis, H Gurney, K Pittman, D Goldstein, P Mainwaring; *Canada*—J Knox, S Ades, T Cheng, S Hotte, Y-J Ko, M MacKenzie, S North; *France*—A Caty, F Rolland, C Chevreau, B Duclos, S Negrier; *Germany*—J Gschwend, P Albers, L Bergmann, J Beck; *Italy*—E Bajetta, R Passalacqua, C Sternberg, F Boccardo, G Carteni, P F Conte; *Japan*—N Shinohara, N Tsuchiya, H Akaza, H Fujimoto, M Niwakawa, H Uemura, H Kanayama, M Eto, Y Sumiyoshi, T Tsukamoto, M Usami, A Terai, Y Hamamoto, M Maruoka; *Netherlands*—S Osanto, C Van Herpen, F Van Den Eertwegh, G Groenewegen; *Poland*—C Szczylik, J Pikiel, A Pluzanska, R Zdrojowy; *Spain*—F del Muro, M Climent, D Castellano, E Calvo, P Maroto; *USA*—N Gabrail, L Appleman, D George, J Hamm, A Hussain, J Hajdenberg, N Vogelzang, T Logan, J Beck, K Rathmell, P Lara, A Dudek, U Vaishampayan, M Gordon, T Anderson, M Danso, W Berry, R Gersh, G Guzley, D Loesch, D Schlossman, D Smith.

Appendix A. Alzheimer's Disease Cooperative Study participants

Arizona Health Sciences Center: Geoffrey Ahern, M.D., Ph.D., Carol Kells, M.S., Keith Burton; Barrow Neurology Clinic: Arthur Schwartz, Ph.D., Charles Echols, M.D., Marci Zomok, R.N., Lauren Dawson, Ph.D.

Brown University: Brian Ott, M.D., Melissa Clemens, B.S., Janet Grace, Ph.D.; Geriatric Med. Res. Group: Sultan Darvesh, M.D., Ph.D., Joanne Cross, R.N., C.C.R.C., Glenda Sherwood; Kansas University, Kansas City: Grisel J. Lopez, M.D., Phyllis Switzer; Mayo Clinic, Jacksonville: Neill Graff-Radford, M.D., Francine Parfitt, M.S.C., C.C.R.C., Lauren M. Makarov, B.S.; Mayo Clinic, Rochester: David S. Knopman, M.D., Bradley Boeve, M.D., Nancy Haukom, R.N., B.A.N., Martha Mandarino, B.A., Diane Mullinax, B.A., Ronald Petersen, M.D., Ph.D.; McGill Centre for Studies in Aging: Serge Gauthier, M.D., Donna Amyot, R.N., MScA, C.C.R.C.; Nathan Kline Institute for Psychiatric Research: Nunzio Pomara, M.D., Corazon de la Pena, M.D.; Northwestern University: M. Marsel Mesulam, M.D., Laura Herzog, M.A.; Oregon Health Sciences University: Jeffrey Kaye, M.D., Joyce Lear, R.N., M.N., Sarah Berman, B.A., Kathy Wild, Ph.D.; Sunnybrook Health Sciences: Sandra Black, M.D., Joanne Lawrence, Maureen Evans; U.B.C. Clinic for Alzheimer's Disease: Howard Feldman, M.D., Valarie O'Neill, R.N., Karen Gilchrist, B.A.; University Hospitals of Cleveland: David Geldmacher, M.D., Cony Santillan, M.D., Parisa Talea, B.S., Marianne Sanders, R.N.; University of California, Los Angeles: Jeffrey Cummings, M.D., Donna L. Masterman, M.D., Michele F. Carter, R.N., Nicole Bennett, MA, Laurie Berndt, B.S.

MEND-CABG II Investigators: G. Walterbusch, Dortmund, Germany, St-Johannes-Hospital (103); F. Isgro, Ludwigshafen, Germany, Klinikum Ludwigshafen (96); C. Brown, Saint John, New Brunswick, Canada, Saint John Regional Hospital (84); R. Cherukuri, Saginaw, Michigan, St Mary's of Michigan (84); F. Malik, Charleston, West Virginia, CAMC Health Education & Research Institute Inc (82); A. Zacharias, Toledo, Ohio, St Vincent Mercy Medical Center (81); J. Ladowski, Fort Wayne, Indiana, Indiana Ohio Heart Group (80); N. Baumgartner, Saginaw, Michigan, Covenant Health Care (76); A. Lamy, Hamilton, Ontario, Canada, Hamilton General Hospital (72); S. Boyce, Washington, DC, Washington Hospital Center (71); T. Yau, Toronto, Ontario, Canada, Toronto General Hospital (70); H. Warnecke, Bad Rothenfelde, Germany, Schuechtermann Klinik (70); R. Holmes, Saginaw, Michigan, Bay Regional Medical Center (60); M. Carrier, Montreal, Quebec, Canada, Montreal Heart Institute (51); F. Dagenais, Quebec City, Quebec, Canada, Hospital Laval (51); A. Bouchard, Birmingham, Alabama, Baptist Medical Center (48); C. Roberts, Winchester, Virginia, Winchester Medical Center (47); J. Rich, Norfolk, Virginia, Sentara Norfolk General Hospital (45); S. Kwan, Montgomery, Alabama, Jackson Hospital (45); W. Killinger, Raleigh, North Carolina, Wake Medical Center (44); L. Collazo, Falls Church, Virginia, INOVA Fairfax Hospital (40); N. Kouchoukos, St Louis, Missouri, Missouri Baptist Medical Center (40); A. Hoeft, Bonn, Germany, Universitaetsklilnikum Bonn (40); C. Randleman, Birmingham, Alabama,

STRADIVARIUS Investigators: Executive Committee: Steven Nissen (Chair), Cleveland Clinic, Cleveland, Ohio; Christopher Cannon, Brigham & Women's Hospital, Boston, Massachusetts; John Deanfield, University College, London, England; Jean-Pierre Després, Laval Hospital Research Center, Quebec City, Quebec, Canada; Christophe Gaudin (nonvoting) and Bernard Job (nonvoting), sanofi-aventis, Pari France; John Kastelein, Academic Medical Center, Amsterdam, the Netherlands; Josep Rodès-Cabau, Laval Hospital, Quebec City, Quebec, Canada; Steven Steinhubl, MD, University of Kentucky, Lexington. **Data and Safety Monitoring Board:** Robert Frye (Chair), Mayo Clinic, Rochester, Minnesota; Robert Harrington and Kerry Lee, Duke University, Durham, North Carolina; Marc Buyse, International Drug Development Institute, Belgium. **Investigators (*indicates sites that were activated but no screening of patients was performed during the study): Australia:** *New South Wales*: John Hunter Hospital, New Lambton (G. Bellamy, MD); Royal Prince Alfred Hospital, Camperdown (M. Adams, PR); *Queensland*: Core Research Group, Milton (D. Colquhoun, PR); *Victoria*: Monash Medical Centre Cardiovascu Research, Clayton (J. Cameron, PR). **Belgium:** Cliniques Universitaires Saint Luc Service de Pharmacie, Bruxelles (J. Renkin, PR); Ziekenh Oost-Limburg Dienst Apotheek, Genk (M. Vrolix, MD); UZ Antwerpen Dienst Apotheek, Edegem (C. Vrints, PR); OLV Ziekenhuis Dienst

Fig 1. Examples of paragraphs in journal articles containing investigator names

features. We found that the predicted accuracy of the SVM does not change much beyond including the 2nd order contextual features for each word, that is, the features from the two neighboring words on either side. Therefore, we choose to include only 2nd order contextual features. Each word feature vector then has an additional 16 x 4 features from the left two and right two neighboring words. A total of 80 features are extracted for each individual word.

Basic Features

Upper case – word starting with a capital letter
Digit – a digit pattern or word with digits
Punctuation – semi-colon, hyphen, dash, comma before or after a word
Name initial pattern – capital character followed by a period

Dictionary Features

First name – word contained in first name dictionary
Last name – word contained in last name dictionary
Affiliation name – word contained in affiliation name key word dictionary
Country name – word in country name dictionary
State name – word in a list of US states and Canadian provinces
Title – word indicating a title such as BS, MD, PhD, etc.

Contextual Features

For current word, the basic and dictionary features extracted from neighboring words

Fig 2. Features extracted for a word

3. STACKED SEQUENTIAL LEARNING

Stacked learning is a meta-learning algorithm that boosts a base learner by taking into account the predicted labels of surrounding tokens. First the base learner is trained, and then the predictions from the base learner are added into the features. Consequently, the data with extended features are used for retraining the learner and reclassification. Based on the fact that the labels of the neighbors can be a clue to predicting the tag of the current token, we believe that adding neighboring token labels as features exploits contextual information present in a sequence and helps improve the base learner.

We choose an SVM classifier as the base learner in our stacked sequential learning due to its ability to accommodate high feature dimensionality and correlated features. The learning and inference steps in the proposed method are outlined in Figure 3. In actual recognition, the ground truth labels are never known. Only the predicted labels from the first SVM

```
Parameter: left and right order $k$
Learning: Given training samples and a base learner SVM,
    1. Train SVM classifier using the initial features and get SVM model $L_0$;
    2. Obtain the predicted labels of the training samples;
    3. Extend the feature vector of each training sample by adding the predictions of $k$
       left and right neighbors;
    4. Retrain SVM classifier and get SVM model $L_1$.

Inference: Given test samples,
    1. Classify the test set with the initial features using $L_0$;
    2. Extend the feature vector of each test sample by adding the predictions of $k$ left
       and right neighbors;
    3. Reclassify the extended test set using $L_1$.
```

Fig. 3. The outline of stacked sequential learning algorithm

SVM classification are available. To avoid the training sample bias problem and to ensure the consistency between learning and inference, at the second SVM training, it is necessary to use the predicted labels of neighboring tokens instead of the true labels as current token features[12]. In addition, to take full advantage of context in our task, the predicted labels of both right and left tokens are used to augment features. This is different from conventional sequential learning such as HMM where current state is only conditional on the previous state.

4. EXPERIMENTS

We collected 5 investigator name paragraphs containing 2087 words as training samples. 11997 words from an additional 8 text paragraphs with long lists of investigator names are used as testing samples. We manually labeled each word in both training and test data. Processing the samples begins with 80 initial features without neighboring word label features included.

We use LibSVM [13], an SVM library developed at National Taiwan University, to implement our word classification. We adopted Radial Basis Function (RBF) as the kernel function where the two parameters, C (penalty parameter of the errors) and γ (RBF parameter), are selected through exhaustive grid-search using cross-validation on the training samples [13].

Following the steps described in Figure 3, at the learning stage we train the SVM classifier twice: at the first level, on the training words with the initial 80 features; at the second level, on the same training samples with features augmented by the neighbors' predictions of the SVM classifier learned from the first level. At the inference phase, we first classify the test words and obtain the predictions of each word, and then the same number of label features from the neighboring words as in the training step are added into the feature set of each word. The test words are finally reclassified using the SVM model trained at the second level.

To observe the gain from stacked learning, we repeat the same scheme by varying the number of neighboring words. The SVM word classification results with different numbers of label features added are shown in Table 1. Left *k* and right *k* words mean *k* neighboring words on either side of the current target word. The classification errors gradually decrease after adding the predictions of surrounding words, which indicates that the stacked sequential learning helps improve accuracy. Out of 11997 test words, the errors reduce from 492 when no label features are used to 423 after 12 label features from the 6 neighbors on either side are included in the current word features. The corresponding accuracy increases from 95.89% to 96.47%. Figure 4 shows that the SVM word classification improves with the increase in the number of neighboring word labels added as features of each target word. As compared to the SVM classifier (a non-sequential learner), the stacked sequential learning enhances the SVM's ability of learning a sequence, thereby improving the accuracy of word classification by exploiting contextual information present among tokens. On the other hand, the computation time will be doubled due to the SVM classification done twice on the test set: first on the original 80 features to get the predicted label of each word, and second on the extended features to obtain the final word predictions. The computation will not significantly increase as long as the number of label features added is far below the initial feature dimensionality.

Table 1. SVM word classification results at different numbers of word labels added as features

Errors Features	Errors (Total number of words: 11997)	Accuracy
Without neighboring word label features, 80 features	492	95.89%
With 1 left and 1 right word label features added, 80 + 2 features	453	96.22%
With 2 left and 2 right word label features added, 80 + 4 features	436	96.37%
With 3 left and 3 right word label features added, 80 + 6 features	435	96.37%
With 4 left and 4 right word label features added, 80 + 8 features	434	96.38%
With 5 left and 5 right word label features added, 80 + 10 features	432	96.40%
With 6 left and 6 right word label features added, 80 + 12 features	423	96.47%

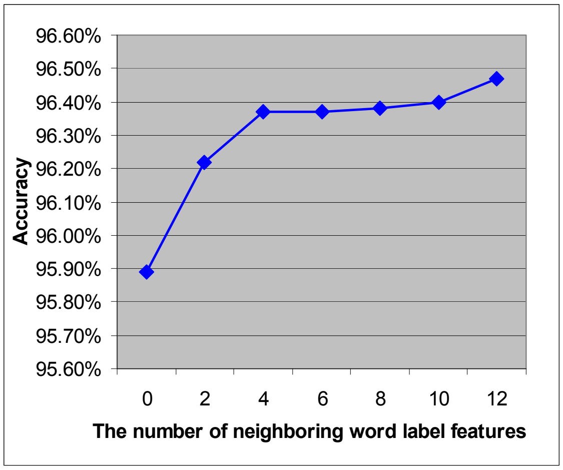


Fig. 4. The SVM word classification accuracy vs. the number of neighboring word labels added as target word features

Tables 2 and 3 show the confusion matrices for the two situations: without label features and with 12 labels from 6 left and 6 right words added as features. In the two tables, type I error is the error of rejecting a hypothesis when it is true, while type II error is the error of failing to reject a hypothesis when it is false. Both errors imply that a sample is placed in a different class other than its true category. For example, in Table 2, the diagonal elements are the numbers of words in each class whose labels are correctly predicted. As shown, out of 4101 First Name words, 3888 are correctly recognized, 24 are misclassified as Last Names and 189 as Others, introducing 213 type I errors. Similarly, 53 Last Names and 63 Other words are misclassified as First Names, producing 116 type II errors. After 12 label features are added, most type I and type II errors decrease. The confusion between First and Last Names is not as important as the confusion between these names and Other words due to the fact that a full name can still be retrieved. We notice a decrease in the type I errors in Other words classified as First or Last Names, as well as in the type II errors in either names classified as Other words. Note that the number of correctly identified First Names increases from 3888 to 3910 and the number of Last Names from 3862 to 3878, respectively, which is very important for improving investigator name retrieval.

Table 2. SVM word classification errors with no neighboring word label features

	First Name	Last Name	Other	Type I Error
First Name	3888	24	189	213
Last Name	53	3862	91	144
Other	63	72	3755	135
Type II Error	116	96	280	492

Table 3. SVM word classification errors with 6 left and 6 right word labels added as features

	First Name	Last Name	Other	Type I Error
First Name	3910	55	136	191
Last Name	41	3878	87	128
Other	53	51	3876	104
Type II Error	94	106	223	423

5. CONCLUSION

In this paper, we describe a stacked sequential learning approach to boost the performance of an SVM classifier for investigator name recognition from online medical articles. Identifying first and last names from a paragraph with special grammatical structure is a task different from typical NER applications. Stacked sequential learning exploits contextual information by including the neighboring word predictions as features and enhances the sequential learning ability of a base learner. Our experiments show that stacked sequential learning improves the performance of an SVM classifier for investigator name recognition when contextual information is present in a sequence of named entities.

It may be noted that stacked sequential learning can have more than two levels with each level using a different learner. Future work is planned on applying multiple stacked sequential learning to the problem of tagging.

6. ACKNOWLEDGEMENTS

We thank Loc Tran for downloading the journal article HTML files and Dr. In-Cheol Kim for providing initial samples of investigator names. This research is supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine, and Lister Hill National Center for Biomedical Communications.

REFERENCES

[1] Nadeau D. and Satoshi S., "A survey of named entity recognition and classification," *Linguisticae Investigations*, 30(1): 3-26 (2007).

[2] Bikel D. M., Schwartz R. L., and Weischedel R. M., "An algorithm that learns what's in a name," *Machine Learning*, 34: 211-231 (1999).

[3] McCallum A., Freitag D., and Pereira F., "Maximum entropy models for information extraction and segmentation," *Proc. of the 17th International Conference on Machine Learning*, 591-598 (2000).

[4] Lafferty J., McCallum A., and Pereira F., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," *Proc. of the 18th International Conference on Machine Learning*, 282-289 (2001).

[5] McCallum A. and Li W., "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons," *Proc. of 7th Conference on Natural Language Learning (CoNLL-2003)*, 4: 188-191(2003).

[6] Carvalho V. R. and Cohen W. W., "Learning to extract signature and reply lines from email," *Proc. of the Conference on Email and Anti-Spam 2004*, Mountain View, California, 2004.

[7] Nguyen N. and Guo Y., "Comparisons of sequence labeling algorithms and extensions," *Proc. of the 24th International Conference on Machine Learning*, 681-688 (2007).

[8] Downey D., Broadhead M., and Etzioni O., "Locating complex named entities from Web text," *Proc. of International Joint Conferences on Artificial Intelligence (IJCAI)*, 2733-2739 (2007).

[9] Minkov E., Wang R. C., and Cohen W. W., "Extracting personal names from email: applying named entity recognition to informal text," *Proc. of the Conference of Human Language Technology and Empirical Methods in Natural Language Processing (HLT)*, 443-450 (2005).

[10] Zou J., Le D., Thoma G.R., "Combining DOM tree and geometric layout analysis for online medical journal article segmentation," *Proc. Joint Conference on Digital Libraries*, 119-128 (2006).

[11] Zou J., Le D., Thoma G.R., "Online medical journal article layout analysis," *Proc. SPIE-IS&T Electronic Imaging 2007, 14th Document Recognition and Retrieval Conference* 6500, 1-12 (2007).

[12] Cohen W. W. and Carvalho V. R., "Stacked Sequential Learning," *Proc. of the 19th International Joint Conferences on Artificial Intelligence (IJCAI)*, 671-676(2005).

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