Named Entity Recognition in Affiliations of Biomedical Articles Using Statistics and HMM Classifiers

Jongwoo Kim and George R. Thoma

Abstract—**This paper proposes an automated algorithm that extracts authors' information from affiliations in biomedical journal articles in MEDLINE® citations. The algorithm collects words from an affiliation, estimates features of each word, and uses a supervised machine-learning algorithm called Hidden Markov Model (HMM) and heuristics rules to identify the words as one of seven labels such as city, state, country, etc. Eleven sets of word lists are collected to train and test the algorithm from 1,767 training data set. Each set contains collections of words ranging from 100 to 44,000. Experimental results of the proposed algorithms using a testing set of 1,022 affiliations show 94.23% and 93.44% accuracy.**

Keyword- Named Entity Recognition, HMM, Heuristic Rule, MEDLINE

I. Introduction

THE U.S. National Library of Medicine (NLM) maintains the MEDLINE database, a bibliographic database containing over 22 million citations related to the biomedical journal literature [1]. Each citation includes fifty-one different fields for each record. NLM receives journal article citations in XML format directly from journal publishers and adds additional fields to the record which are provided by journal article indexers. The number of citations in MEDLINE is rapidly increasing every year. NLM collects statistics such as the number of citations for each publication year, the number of citations of total year, the number of citations with authors in total year, etc. However, there are no detailed statistics such as, the number of citations published per each country each year, the number of citations per each organization each year, or the number of citations that received grants from NIH per each country each year, etc. In addition, there is no citation field for country, organization, etc. in the existing citations. Therefore, extraction of such fields from authors' affiliations is critical for the collection of detailed statistics.

There are several studies on extracting authors' information from texts. Robinson et al. extracted affiliations from free texts [2] and Kim et al. [3] extracted affiliations from journal articles. Further studies have been done to extract information from the authors' affiliations. Yu et al. [4] used word dictionaries and regular expression to extract three labels (institution, country and email address). Jonnalagadda et al. [5] used rules and word dictionaries to extract eight different labels. Torii et al. [6] used a HMM package to label eight different labels for words in affiliations. However, the papers did not address the following issues. First, labeling a word that contains two or three labels (no separators between labels). Second, labeling affiliations without country name. Third, labeling affiliations that contain more than two organization names. Fourth, usage of label orders and relationship between labels. Fifth, usage of probabilities of each word for each label.

Therefore, we propose a prototype of an automatic algorithm handling the above issues to classify words in affiliations into seven different labels to collect detailed statistics. HMM, statistics, and heuristic rules are adapted for the algorithms.

The remainder of this paper is organized as follows. Section II describes authors' affiliations in articles. The details of our methods are presented in Section III and IV. We discuss experimental results in Section V, and show conclusions in Section VI.

II. Authors' Affiliations

Words in authors' affiliations can be categorized by two groups, organization and geographic terms. Organization terms include department, school, university, institute, etc. Geographic terms include city, state/province, postal code, and country. Email can belong to either one or two groups since it can contain country and organization names. There are several types of affiliations based on the orders of the term words. Some affiliations show just organization names and others show the full address (mailing address) of their organizations. Our final goal is to label affiliation words into nine different labels (Department, School, University, City, State/Province, Postal Code, Country, Email, and Other). In this paper, as a preliminary work, we label affiliation words into seven labels (University, City, State/Province, Postal Code, Country, Email, and Other). In private organizations, "University" means company names, "School" means institutes or centers that belongs to the companies, and "Department" means departments or divisions that belongs to the institutes or center.

Table I shows some examples of affiliations. In the table, PMID (PubMed Unique Identifier) is a unique reference number for the MEDLINE citations. "Un" means "University", "Ci" means "City", "St" means "State", "Co" means "Country", "Po" means "Postal Code", "Em" means "Email", and "Ot" means "Other". Some affiliations have a company name only (Type 1), some have full address including street name (Type 2), some have a country name

J. Kim is with the National Library of Medicine, Bethesda, MD 20893, USA (corresponding author to provide phone: 301-435-3227; fax 1 301 402-0341; e-mail: jongkim@mail.nih.gov).

G. R. Thoma is with the National Library of Medicine, Bethesda, MD 20893, USA

(Type 3), some do not have a country name (Type 4), and others have an email address (Type 5). We define the type based on the label orders and use it to develop the algorithm.

TABLE I
AFFILIATIONS IN ARTICLES IN MEDLINE.

Type	Explanation/Examples	Label Order	PMID
1	bioMerieux, Inc.	Un	23002511
2	Faculty of Kinesiology, University of Calgary, 2500 University Drive NW, Calgary, Alberta, Canada T2N 1N4.	Ot, Un, Ci, St, Co, Po	23000101
3	Department of Urology, School of Medicine, University of Gaziantep, 27310 Gaziantep, Turkey.	Ot, Ot, Un, Po, Ci, Co	23001641
4	Department of Psychology, University of Houston.	Ot, Un	23000106
5	Department of Physics, The Ohio State University, Columbus Ohio 43210, USA.	Ot, Un, Ci, St, Po, Co	23002754
6	Department of Psychology, School of Life and Medical Sciences, University of Hertfordshire, UK. k.laws@herts.ac.uk	Ot, Ot, Un, Co,Em	23001963
7	Zaklad Medycyny Nuklearnej Pomorskiego Uniwersytetu Medycznego w Szczecinie ul. Unii Lubelskiej 1, 71-252 Szczecin.	Ot, Un, Po, Ci	23002662
8	Division of Gastroenterology, Hepatology, and Nutrition, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA.	Ot, Un, Ci, St,, Co	23000233
9	Department of Global Safety Pharmacology, Department of Pharmacokinetics, Dynamics & Metabolism, and Neuroscience, Pfizer Global Research and Development Eastern Point Road, Groton, CT 06340, USA. anthony.fossa@icardiac.com	Ot, Un, Ci, Po, Co, Em	23000177

III. ALGORITHM WORKFLOW

The workflow of the proposed algorithm is as follows. First, replace words (names) with standardized words using a dictionary. Second, divide an affiliation into words using several separators. Third, divide a word with more than two labels into several words using geographic information (city, state/province, country, postal code, etc.) in collected word lists. Fourth, assign a possibility value for each word for each label using collected word lists and statistics. Fifth, classify each word as one of the seven labels using HMMs. More detailed information of each step is shown in the next section.

IV. PROPOSED APPROACH

We use the following word lists and algorithms for labeling words in affiliations.

A. *Word Normalization*

There are several abbreviated (or non-standard) words in affiliations. We first standardize the words. Table II shows some examples collected from a training set of 1,767 affiliations in MEDLINE. For example, authors use several ways to write the country name "China" as shown in the second row. All non-standard words are replaced with our own standard words and abbreviated names are replaced with full names also. For example, when the organization name "NIH" is replaced with "National Institutes of Health", it becomes clear that "NIH" is an institution name. 155 words are collected related to city, country, organization name, and other words for the list.

TABLE II
LIST OF WORDS FOR STANDARDIZATION

Standard Word	Non-Standard Word
China	P R China, People's Republic of China, PR of China, etc.
Germany	Deutschland, Federal Republic of Germany, F.R.G, etc.
Korea	Republic of Korea, South Korea
National Institutes of Health	NIH
Technische Universität Berlin	TU Berlin

B. *Postal Code Detection*

Every country has their postal codes formats. We search the codes of several countries using Google search engine [7], collect the codes for 121 countries, and save them as Regular Expression [8] formats. Table III shows the formats of some countries.

TABLE III
LIST OF POSTAL CODE FORMATS OF COUNTRIES

Country Name	Postal Code (Regular Expression)
Austria	\\b((A-|)[0-9O]{4})\\b
Brazil	\\b(([0-9O]{5}|[0-9O]{2}[.][0-9O]{3})[-][0-9|O]{3})\\b
India	\\b(([0-9O]{6})|([0-9O]{3}[ ][0-9O]{3}))\\b
USA	\\b([A-Z]{2}(|)[0-9O]{5}[-][0-9O]{4})\\b
Vietnam	\\b([0-9O]{6})\\b

C. *City, Region, and Country names Detection*

A list of city, state, and country names are necessary to recognize them from affiliations. First, we collect names from affiliations in MEDLINE. In addition, we use Google search engine [7] to collect more information. There are about 44,000 names in the list. Table IV shows examples of some of the data.

TABLE IV
LIST OF CITY, REGION, AND COUNTRY NAMES

Country	Region (State/Province)	City/Town
Australia	New South Wales	Sydney
Canada	Quebec	Montreal
China	Shaanxi	Xi'an
Germany	Nordrhein-Westfalen	Düsseldorf
Germany	North Rhine-Westphalia	Dusseldorf
Philippines	Sorsogon	San Juan
South Africa	Eastern Cape	Grahamstown
Spain	Pontevedra	Vigo
SAUDI Arabia	Makkah	Thuwal
USA	Maryland	Bethesda

D. Organization Name Words

Organization names are categorized into three labels (Department, School, and University levels). It is clear when affiliations are from universities. However, it is hard to categorize affiliations from companies, laboratories, etc. Therefore, we search for affiliations in the training set, classify them into the three labels using the Google search engine, and collect the words related to the three labels as shown in Table V. Among them, several words are used in more than two labels. Table VI shows the probabilities of the three labels for some words in Table V. For example, "Center" is used for University (University level) 51 times, School 31 times, and Department 15 times. The probabilities are used to classify organization names at the University level.

TABLE V
LIST OF WORDS FOR ORGANIZATION NAMES

Academy	Fachbereich	Laboratorios
Aquarium	Katedra	Laboratorie
Agence	Division	Laboratoria
Agency	Engineering	Laboratorium
Association	Faculty	Laboratory
Branch	Faculte	Laboratório
Bureau	Faculté	Laboratorio
Campus	Fakultät	Laboratoire
Center	Faculdade	Library
Centre	Facultad	Limited
Centro	Faultad	LLC
Központ	Facoltà	Ministry
Centrum	Kar	Museum
BioCenter	UFR	Organization
Hemocentro	Wydział	Organisation
Herzzentrum	Wydzial	Pharmaceutical
Clinic	Foundation	Pharma
Clinics	Fundación	Pharmacal
Clinico	Fund	Pharmaceutica
Clínico	Group	Pty
Clínica	Hospice	School
Klinik	Hospices	Services
Klinika	Hospital	Society
Kliniki	Hospitals	Trust
Poliklinigi	Hôpital	University
College	Hôpitaux	Universitat
Collegium	Hospitalier	Universiti
Hochschule	Klinikum	Université,
Charities	Ziekenhuis	Universite
Company	Ospedale	Università
Commission	Ospedaliera	Universitaria
Committee	Ospedaliero	Universität
Corporation	Hastanesi	Universiteit
Council	Incorporated	Universidad
Consiglio	INC	Universidade
Department	Inc	Universitaire
Départment	Inc.	Universitätsspital
Departments	Institutes	Universitätsklinikum
Departament	Institution	Universitätskliniken
Département	Institute	Universitätsmedizin
Departement	Institut	Uniwersytetu
Dipartimento	Instituto	Uniwersytet
Departamento	Instytut	Nationale Supérieure
Deparment	Institutet	Egyetem
Dept.	Istituto	Tudományegyetem
Dept	Intézet	Unit
Dpto	Laboratories	Unité

TABLE VI
PROBABILITIES OF WORDS FOR UNIVERSITY, SCHOOL, AND DEPARTMENT LABEL

Affiliation Words	Prob. of University	Prob. of School	Prob. of Department
Center, Centre, Centro, Központ, etc.	0.5258	0.3196	0.1546
Department, Départment, Département, Dipartimento, etc.	0.0043	0.0239	0.9717
Faculty, Faculte, Faultad, Facoltà, etc.	0.0625	0.8906	0.0469
Hospital, Hôpital, Hôpitaux, Klinikum, Hastanesi, etc.	0.7383	0.2523	0.0093
Institute, Institution, Institut, Intézet, etc.	0.4779	0.4412	0.081
Laboratory, Laboratorios, Laboratorium, Laboratoire, etc.	0.0833	0.3000	0.6167
University, Universitat, Universitaria, Uniwersytet, etc.	0.9795	0.0154	0.0051

E. Other Words

There are words such as road name, building number, subdivision name, etc. in affiliations. These words are labeled as Other. Table VII shows some words collected for the Other label.

TABLE VII
LIST OF WORDS FOR OTHER LABEL

Other Word Category	Words
Road	Avenue, Avenida, Freeway, Route, Street,
Sub division	Ro, Ku, Gu, etc.
Building	Suit, Building,
P.O. Box	P.O.Box, PO Box, POB, Private Bag, etc.

F. Email Address

There are several Regular expression formats to recognize email addresses in affiliations. Among them, we use the following format of Regular Expression [9].

```
"([a-zA-Z0-9_\\-\\.]+)@((\\[[0-9]{1,3}\\.[0-9]{1,3}\\.[0-
9]{1,3}\\.)|(([a-zA-Z0-9\\-]+\\.)+))([a-zA-Z]{2,4}|[0-
9]{1,3})(\\]?)"
```

G. Abbreviate Organization Name Detection and Removal

Some authors write their organization names twice in affiliations; including full and abbreviated names. Table VIII shows some examples. Abbreviated names are usually enclosed in parentheses as shown in Type 1 to 5. However, all of them are not duplicated names. Type 6 shows that state name VA (province of Varese) is enclosed in parentheses. The following two methods are used to remove the abbreviated names before classifying words in affiliations.

TABLE VIII

AFFILIATIONS HAVING TWO SAME ORGANIZATION NAMES

Type	Affiliation with full and abbreviated organization names
1	Burn Research Center (BRC), Shahid Motahari Burns Hospital, Tehran University of Medical Science, Tehran, Iran. bsobooti@tums.ac.ir
2	Instituto de Biología Molecular y Celular (IBMC), Miguel Hernández University, 03202, Elche, Spain.
3	National Institute of Advanced Industrial Science and Technology (AIST), Umezono, Tsukuba 305-8568, Japan.
4	Laboratoire d'ergonomie et d'épidemiologie en santé au travail (LEEST), LUNAM Université, Université d'Angers, LEEST-UA InVS, Angers, France. celine.serazin@univ-angers.fr
5	Center of Calcium and Bone Research (COCAB), Mahidol University, Bangkok, Thailand.
6	U.O. Cardiologia-Emodinamica Istituto Clinico Humanitas Mater Domini, Via Gerenzano 2, Castellanza (VA), Italy. alielasi@hotmail.com

Method 1 works for the Types 1, 2, 3, and 4 and Method 2 works for Type 5. However, no method is working for the Type 6 case.

Method 1: (see Type 1 in Table VIII)

Step 1, replace several words (e.g., “d'é” to “E”, “d'É” to “E”, etc.) in an affiliation.

Step 2, find a name (“BRC”) in a parenthesis or with all uppercase characters.

Step 3, collect all words using space, comma, semi-colon, colon, and parenthesis as separators.

Step 4, make a string (“BRCBSMBHTUMSTI”) using the first uppercase character in each word.

skip a word if the first character in the word is not uppercase character.

Step 5, remove the name (“BRC”) from the affiliation if it is found in the string in Step 4.

Method 2: (see Type 5 in Table VIII)

Step 1, replace several words (e.g., “d'é” to “E”, “d'É” to “E”, etc.) in an affiliation.

Step 2, find a name (“COCAB”) in a parenthesis or with all uppercase characters.

Step 3, collect words using comma, semi-colon, colon, and parenthesis as separators.

For each word (having more characters than the name.)

Step 4, search each character in the name (COCAB) in the word (“Center of Calcium and Bone Research”).

Step 5, remove the name from the affiliation if all characters in the name are found in the same order in the word and 75% of matched characters in the word are uppercase.

H. *Word Separation*

Seven separators (“,” “;” “:”, “(“, “)”, “[“, “]”) are used to separate words in an affiliation. The separators works well for most affiliations. However, many authors use “white space” as a separator frequently. Type 5 in Table I uses a “white space” separator between city, state, and postal code “Columbus Ohio 43210”. Type 7 in Table I also uses a “white space” separator between department, school, and university. This causes labeling errors or increases computation time to separate the words in affiliations. In the case of Type 7 in Table I “Zaklad Medycyny Nuklearnej Pomorskiego Uniwersytetu Medycznego w Szczecinie ul. Unii Lubelskiej 1, 71-252 Szczecin”, since there is no separator between department and university as is written in the Polish language, it causes word separation and University (organization name) labeling errors. This also creates errors in the Postal Code and City labels. Therefore, the following steps are used to separate words for accurate labeling.

```
Step 1. Divide words ($w_i$, where $i$=1 to $n$) in an
   affiliation using the seven separators.

Step 2. Search email.
  If $w_i$ contains an email,
    Divide a word $w_i$ into $w_{ij}$, where $j$=1 to $k$,
      using “white space”.
    Replace $w_i$ with $w_{ij}$, where $j$=1 to $k$.
    Update $n$=$n$+$k$-1.
    Stop.
  End If

Step 3. Search country name using Table IV.
  If $w_{n-1}$ or $w_n$ is not country name.
    For $i$=$n$ to $n$-1
      Divide a word $w_i$ into $w_{ij}$, where $j$=1 to $m$,
        using “white space” if no word found in Table V.
      If country name found in $w_{ij}$,
        Replace $w_i$ with $w_{ij}$, where $j$=1 to $m$.
        Update $n$=$n$+$m$-1.
        Stop.
      End If.
    End For
  End If.

Step 4. Search city and region names and postal code
   using Tables II, III and IV.
  If $w_i$ is not city, region, or postal code.
    For $i$=$n$ to 1
      Divide a word $w_i$ into $w_{ij}$, where $j$=1 to $p$
        using “white space” if no word found in Table V.
      If one of the labels found in $w_{ij}$,
        Replace $w_i$ with $w_{ij}$, where $j$=1 to $p$.
        Update $n$=$n$+$p$-1.
        Stop.
      End If.
    End For
  End If
```

I. *Adjust Possibilities of Labels*

Heuristic rules are used to adjust possibilities of labels. All of the rules and ratios are based on the statistics obtained from the training set. Table IX shows some of the rules. Rule 1 means if a word (w_i) does not have any clue for City, but the previous word (w_{i-1}) has a possibility for University and the next word (w_{i+1}) has a possibility for State, the word (w_i) has 98% of possibility for City.

TABLE IX
HEURISTIC RULES FOR ADJUSTING POSSIBILITY OF A LABEL

Rule	Condition
1	*If* $P_{University}(w_{i-1}) > 0$, $P_{City}(w_i) = 0$, and $P_{State}(w_{i+1}) > 0$, $Pcity(w_i)$=0.98.
2	If $P_{University}(w_{i-1}) > 0$, $P_{City}(w_i) = 0$, and $P_{Postal\ Code}(w_{i+1}) > 0$, $Pcity(w_i)$=0.80
3	If $P_{University}(w_{i-1}) > 0$ and $P_{University}(w_i) > 0$, $P_{University}(w_i) = P_{University}(w_i) \times 0.9384$. $P_{University}(w_{i-1}) = P_{University}(w_{i-1}) \times 0.0616$.

J. Hidden Markov Model (HMM)

HMM [10] is used to extract labels from affiliations since it provides stable results in named entity recognition areas. In addition, the Viterbi [11] algorithm is used to finalize labels of words in affiliations from the HMM results. To train HMMs using the training set, we group the training set data by the label order and train HMMs for each group. The following is the complete algorithm procedure.

Step 1. Separate words from an input affiliation.
Step 2. Standardize words using Table II.
Step 3. Assign possibilities of Department, School, and University labels using Tables V and VI.
Step 4. Assign possibilities of City, State, Country, Postal Code, and Email labels using Tables III and IV.
Assign 1.00 if a word is in the tables or meets formatting requirements.
Step 5. Assign possibilities of Other label using Table VII.
Step 6. Adjust possibility of labels using in Table IX.
Step 7. Apply all trained HMMs for the input affiliation and select one HMM (HMM_{final}) that has the highest value.
Step 8. Use the Viterbi algorithm in HMM_{final} to finalize labels of words in the affiliation.

V. EXPERIMENTAL RESULTS

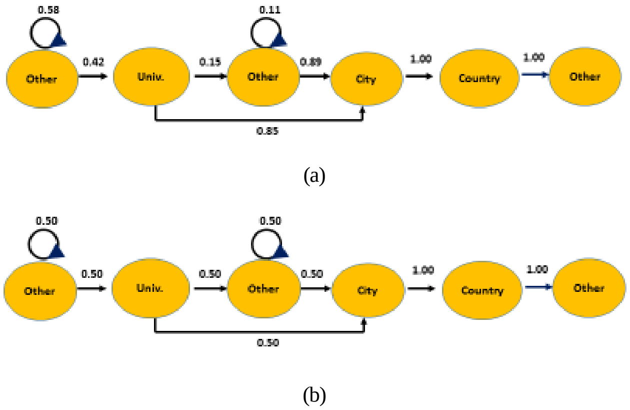
All affiliations in MEDLINE in the ranges from PMID=23,000,000 to 23,004,000 are used in this experiment. Among them, 1,767 PMIDs (from PMID=23,000,000 to 23,002,000) are collected for the training set and 1,022 PMIDs (from 23,002,001 and 23,004.000) for the testing set. Since several PMIDs do not have any information in MEDLINE, training and testing sets do not have a similar amount of data.

To train HMMs, we group the training set data by order of labels first and train HMMs for each group. To optimize the number of training HMMs and number of training data for each HMM, we remove "Other" between labels in the training set data. For example, a training set with data containing "Other, University, Other, City, Other, Postal Code, Other, Country" is assigned to "Other, University, City, Postal Code, Country" group for training.

We have 30 HMMs from the training set. Some HMMs contains a reasonable number of training data. However, other HMMs have less training data. Fifteen HMMs have less than ten training data. Table X shows some of the trained HMMs. The third row in the table shows that the HMM (Other, University, City, Country) has 280 training data and Fig. 1 shows the diagrams of the HMM. Fig. 1(a) shows the HMM from the training data (HMM Trained) and Fig. 1(b) shows the HMM modified from the HMM 1(a) (HMM Modified). i.e., the transition workflows from one label to other labels are the same in the two HMMs. The difference is that Fig. 1(b) has equal transition weights (=1/*k*) when one label can move to *k* different labels.

TABLE X
HMMS TRAINED USING THE TRAINING SET

HMM	**Number of PMIDs used**
Other,University,City,ZipCode,Country	311
Other,University,City,Country	280
Other,University,City,ZipCode,Country,Email	226
Other,University,City,Country,Email	210
Other,University,ZipCode,City,Country,Email	139
Other,University,City,State,Country	109
Other,University,City,State,ZipCode,Country	66
Other,University,Country,Email	37
Other,University,ZipCode,State,Country,Email	1



(a)

(b)

Fig. 1. (a) HMM (HMM Trained) in the third row in Table IX. (b) HMM (HMM Modified) modified from the HMM (a).

Table XI shows the test results. We consider errors when one of the words in an affiliation is miss-labeled by HMMs. The HMMs Trained column shows 94.23% accuracy and HMMs Modified shows 93.35% accuracy. The HMMs Trained shows better performance than the HMMs Modified.

We evaluate the 58 errors from HMMs Trained and categorize them into five as shown in the Table XII. First, the major errors are caused by a word separator problem. As shown in the second row in the table, there are no separators between department and university, and between postal code and city name. It is written in Polish and there is no country name there. In addition, our word lists (Tables V and VI) do not have much information about non-English languages. All these issues make the algorithm more difficult to separate the words. Second, there are not many organization names in our word lists. “IFW Dresden” is the organization (third row). However, the name does not contain any clue word related to organization names. Therefore, “Institute for Integrative Nanosciences” is recognized as the organization name. Third, there is no trained HMM that fits to the input (fourth row). This problem can be resolved by increasing the size of training set data. Fourth, existing HMMs are trained for processing single affiliation. Therefore, the algorithm cannot handle texts with multiple affiliations (fifth row). The last error is the order of labels (sixth row). 93.84% of affiliations in the training set has the order of “Department, School, University”. However, University word “USDA Forest Service” comes first and Department word “Aldo Leopold Wilderness, Research Institute” comes later. Since there is no clear word representing University in “USDA Forest Service”, but the word “Institute” is found in “Aldo Leopold Wilderness, Research Institute”, “Aldo Leopold Wilderness, Research Institute” is labeled as University label. This problem can be resolved by collecting names that belong to University label and use them for labeling.

TABLE XI
PERFORMANCE OF THE PROPOSED HMMS

HMM Mode	HMMs Trained	HMMs Modified
Total	1,022	1,022
True	964	955
False	58	67
Accuracy	94.32%	93.44%

TABLE XII
ERROR ANALYSIS

Error Analysis	Number of PMIDs	Affiliation Example
Word separation error	21	Zaklad Medycyny Nuklearnej Pomorskiego Uniwersytetu Medycznego w Szczecinie ul. Unii Lubelskiej 1, 71-252 Szczecin.
Hard to recognize words in University label	20	Institute for Integrative Nanosciences, IFW Dresden, D-01069 Dresden, Germany. j.zhang@ifw-dresden.de
HMM does not exit	11	Department of Physics, Indian Institute of Technology, Bombay, Powai, Mumbai-400 076, India. supravat@phy.iitb.ac.in
Multiple affiliations	5	Heart Institute, Ha'Emek Hospital, Afula, Israel, affiliated with Rappaport Faculty of Medicine, Haifa, Israel.
Label order problem	1	USDA Forest Service, Rocky Mountain Research Station, Aldo Leopold Wilderness, Research Institute, 790 East Beckwith, Missoula, MT 59801, USA. sean_parks@fs.fed.us

VI. CONCLUSIONS

This paper proposes an automatic algorithm to classify seven different labels from affiliations in biomedical journal articles using statistics, heuristic rules and HMM. We collect seven different word list tables to estimate the possibilities of seven different labels for each word in the author affiliations. We also collect 1,767 affiliations for a training set and 1,022 affiliations for a testing set from MEDLINE.

The proposed module performs relatively well. The results shows 94.23% accuracy from HMMs Trained and 93.35% accuracy from HMMs Modified.

As a future task, we plan to use more data for the training set to handle additional different types of affiliations and collect (international) organization names for more accurate classification. In addition, we will extend the algorithm classifying the nine different labels from affiliations.

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] Chinchor, N, Robinson, P, “MUC-7 named entity task definition”, Proceedings of the 7th Message Understanding Conference, 1997.

[3] Kim J, Le DX, Thoma GR. “Automated Labeling Of Biomedical Online Journal Articles”, SCI 2005. Proc 9th World Multiconference on Systemics, Cybernetics and Informatics; 2005 Vol. 4; Orlando (FL).

[4] Yu, W., Yesupriya, A. et. al., “An Automatic method to generate domain-specific investigator networks using PubMed abstracts’, BMC Medical Informatics and Decision Making, Vol 7, 17, 207.

[5] Jonnalagadda, S. and Topham, p., “NEMO: Extraction and normalization of organization names from PubMed affiliation strings”, Journal of Biomedical Discovery and Collaboration, Vol. 5, 50-75, 2010.

[6] Torii, M.,Wagholikar, K., Kim, D., Liu, H. "Named Entity Recognition in the MEDLINE Affiliation Field: A Step towards Enhanced Maintenance of Researcher Profile Systems" AMIA CRI, pp. 164, 2012.

[7] http://www.google.com.

[8] https://msdn.microsoft.com/en-us/library/hs600312(v=vs.110).aspx.

[9] http://regexlib.com.

[10] Chahramani, Z., “An Introduction to Hidden Markov Models and Bayesian Networks”, International Journal of Pattern Recovnition and Artificial Intelligence, 15 (1), pp. 9-42, 2001.

[11] Viterbi, A. J., Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory* , vol. IT-13, April, 1967, pp. 260-269.