Labeling Author Affiliations in Biomedical Articles Using Markov Model Classifier

Jongwoo Kim, Soyoung Hong, and George R. Thoma

***Abstract*— This paper proposes an automated labeling algorithm that extracts authors' affiliation information (organization, city, country, etc.) from the citations in NLM's MEDLINE® database. Researchers and granting organizations can recognize the most active research organizations or countries in specific fields by comparing the number of publications generated in each organization or country. We are developing a system to collect/show such statistics from MEDLINE. Extraction of the authors' information from affiliations in the citations is the key step to obtaining the statistics. The proposed labeling algorithm divides an affiliation into several pieces and identifies each piece as one of seven labels (authors' affiliation information). We adapt Stanford CoreNLP tool, Markov Model (MM), and Viterbi algorithm for the proposed algorithm. Experimental results of the proposed algorithms show 95.90% accuracy.**

Keyword- MEDLINE, Labeling, Stanford CoreNLP tool, Markov Model, Viterbi, Heuristic Rule

I. Introduction

The U.S. National Library of Medicine (NLM) maintains the MEDLINE database, a bibliographic database containing over twenty-six million citations from the biomedical journal literature [1]. NLM collects statistics such as the total number of citations, the number of citations per year, etc., but not the number of citations published per country each year, number of citations published per organization each year, the number of citations that received grants from U.S. federal government each year, etc. The number of publications can be seen as a key measure of active research in specific fields, and may be useful information for researchers, students, and granting organizations. We are developing a system to collect such statistics and provide them to public users through a website [2]. Every citation in MEDLINE includes fifty-one different fields. However, there are no separate fields for authors' organization name, city, country, etc. Therefore, automatic extraction of these fields from authors' affiliations is critical for the collection of the statistics.

There are several research articles concerned with extracting information from authors' affiliations. Jonnalagadda et al. used word dictionaries and rules to extract eight different labels and normalized ambiguous institution names [3]. Yu et al. used regular expressions and word dictionaries to extract institution, country and email address [4]. Torvik extracted city, state, country, and the longitude and latitude of the location information from affiliations in PubMed using a set of (city, state, and country) and n-grams word list [5]. Bhargava et al. extracted people, places, and organizations from sentences written in English [6]. Torii et al. used a Hidden Markov Model (HMM) package to estimate eight different labels for words in affiliations [7]. Stanford Named Entity Recognizer labels texts into four tags (Location, Person, Organization, Misc) [8]. However, the papers did not address labeling words that contain two or three labels without any separator between labels, usage of probabilities of each word for each label, usage of relationship between labels, etc. There are issues of dividing an affiliation into several pieces before labeling. However, no papers have addressed these issues. In addition, the papers did not separate organization names into levels such as university, school, and department.

An early version of our work was presented in [9]. This work also had issues dividing an affiliation into pieces, resulting in labeling errors. In this paper, we explore labeling algorithms in more depth, especially with respect to these dividing issues. Our proposed labeling algorithm divides an affiliation into pieces using separators (comma, semicolon, etc.), the Stanford CoreNLP tool, and heuristic rules. It then classifies the pieces with seven different labels.

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

II. Authors' Affiliations

There are several types (labels) of words in authors' affiliations. Department, school, and university are used for organization names; city, state/province, postal code, and country are used for geographic locations. Email can have organization name and/or geographic locations. We assign nine labels to affiliation words such as Department, School, University, City, State/Province, Postal Code, Country, Email, and Other. For private organizations, "University" stands for company names, "School" for institutes or centers that belongs to the companies, and "Department" for departments or divisions that belong to the institutes or center. In this paper, we label affiliation words into seven labels (University, City, State/Province, Postal Code, Country, Email, and Other) as preliminary work.

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

S. Hong is with the Montgomery Blair High School, Silver Spring, Maryland, USA

Table I shows some affiliation types. In the table, “Ot” means “Other”, “Un” means “University”, “Ci” means “City”, “St” means “State”, “Co” means “Country”, “Po” means “Postal Code”, and “Em” means “Email”. We define the type based on the label orders and use them to develop the algorithm.

TABLE I
AUTHORS’ AFFILIATIONS IN ARTICLES.

Type	Explanation/Examples	Label Order
1	3E Company.	Un
2	Pomerado Hospital, Palomar Pomerado Health, Poway, CA 92064, USA. david.tam@pph.org	Ot, Un, Ci, Po, Co, Em
3	Department of Pharmaceutics and Pharmaceutical Chemistry, The University of Utah, 421 Wakara Way, Suite 318, Salt Lake City, Utah 84108, USA.	Ot, Un, Ot, Ot, Ci, St, Po, Co

III. DATA LISTS USED FOR LABELING

The following word lists and tables are used for labeling affiliations.

A. Word Normalization

We first standardize several abbreviated, non-standard, or foreign language words in affiliations. Table II shows some examples collected from the training set (see Section V) in MEDLINE. For example, there are several ways of writing the country name “Germany” as shown in the first row. “KU Leuven” is also replaced with “Katholieke Universiteit Leuven” because it becomes clear that “KU Leuven” is a university name. 176 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
Germany	Deutschland, Federal Republic of Germany, F.R.G, etc.
Italy	Italia, Italie, etc.
Katholieke Universiteit Leuven	KU Leuven
National Institutes of Health	NIH

B. City, Region, and Country names

We collect a list of city, state, and country names from affiliations in MEDLINE. Google search engine [10] is also used to collect more such information. The list has about 43,700 names. Table III shows some examples.

TABLE III
LIST OF CITY, REGION, AND COUNTRY NAMES

Country	Region (State/Province)	City/Town
Austria	Wien	Vienna
France	Aquitaine	Arcachon
Philippines	Pangasinan	Alaminos
South Africa	Eastern Cape	Grahamstown
Spain	Pontevedra	Vigo
USA	Maryland	Bethesda

C. Postal Code

Postal code formats for 162 countries are collected using Google search engine and saved as Regular Expression [11] formats. Table IV shows the postal code formats used in some countries.

TABLE IV
LIST OF POSTAL CODE FORMATS OF COUNTRIES

<table>
<tr><th>Country Name</th><th>Postal Code (Regular Expression)</th></tr>
<tr><td>Australia</td><td>\\b(([A-Z]{2}|[A-Z]{3})(|)[0-9|O]{4})\\b</td></tr>
<tr><td>Brazil</td><td>\\b(([0-9O]{5}|[0-9O]{2}[.][0-9O]{3})[-][0-9|O]{3})\\b</td></tr>
<tr><td>Korea</td><td>\\b([0-9|O]{3}[-][0-9|O]{3})\\b</td></tr>
<tr><td>Saudí Arabia</td><td>\\b(([0-9|O]{5}[-][0-9|O]{4})|([0-9|O]{5}))\\b</td></tr>
<tr><td>USA</td><td>\\b([A-Z]{2}(|)[0-9O]{5}[-][0-9O]{4})\\b</td></tr>
</table>

D. Organization Name Words

Affiliations usually have three labels (levels) of organization names such as Department, School, and University. In the case of affiliations for universities, these labels are clear. However, it is hard to find the corresponding three labels in the case of private companies or organizations. Therefore, we classify organization names related to the three labels from affiliations in training set using Google search engine, collect common words as shown in Table V, and estimate probabilities of each word for the three labels. Table VI shows examples of the probabilities. In the case of “Hospital”, 74% is used for University (University level), 25% for School, and 1% for Department. These probabilities are used to classify organization names at the University level.

TABLE V
LIST OF WORDS FOR ORGANIZATION NAMES

Academy	Fachbereich	Laboratorios
Agence	Division	Laboratoria
Association	Faculty	Laboratory
Center	Faculdade	Library
Központ	Facoltà	Ministry
Hemocentro	Wydział	Organisation
Herzzentrum	Wydzial	Pharmaceutical
Kliniki	Hospital	Trust
College	Hôpital	University
Corporation	Hastanesi	Universiteit
Department	Inc	Universitaire
Departement	Institute	Uniwersytetu

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

Affiliation Words	Prob. of University	Prob. of School	Prob. of Department
Hospital, Hôpital, Hôpitaux, , etc.	0.7383	0.2523	0.0093
Department, Départment, etc.	0.0043	0.0239	0.9717
Institute, Institution, Institut, Intézet, etc.	0.4779	0.4412	0.081
University, Universitat, etc.	0.9795	0.0154	0.0051

E. Email Address

We investigate several Regular Expression formats for extracting email addresses in texts and choose the following format [12].

```
"([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})(\\]?)"
```

F. Other Words

Some affiliations contain words related to road, building, subdivision, postal office box in affiliations. These words are labeled as Other. Table VII shows some such words.

TABLE VII
LIST OF WORDS FOR OTHER LABEL

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

IV. PROPOSED ALGORITHMS

The proposed algorithm consists of five major steps. First, replace words (names) with standardized words using a dictionary. Second, separate an affiliation into several pieces using separators, geometric information in the collected word lists (Tables III and IV), and heuristic rules. Third, correct the separation error in the second step using the Stanford CoreNLP tool [13] and heuristic rules. Fourth, estimate possibility values of each piece for all labels using the collected tables and heuristic rules. Fifth, label each piece as one of the seven labels using MMs and Viterbi algorithms. The next sections show more detailed information for each step.

A. Affiliation Separation

We use seven punctuation marks as separators (“,” “;” “:”, “(“, “)”, “[“, “]”) to divide an affiliation into several pieces for labeling. These separators perform well for most affiliation cases. However, many authors do not use separators when writing their affiliations. In the case of “Department of Internal Medicine, University of Iowa, Iowa City IA 52242, USA”, white spaces are used as a separator between city, state, and postal code. In the case of “Department of Biological Sciences University of Arkansas, Fayetteville, AR 72701, USA”, white space is also used as a separator between department and university. This causes serious labeling errors. In addition, it increases computation time to separate the pieces again. The following steps are used to separate city, state, country, and postal code from words without punctuation marks separating them (e.g., “Iowa City IA 52242”).

```
Step 1. Separate words ($w_i$, where $i$ = 1 to $n$) in an
affiliation using the seven separators.
Step 2. Search city, state, country, and postal code using
    Tables III and IV.
    If $w_i$ is not city, region, country or postal code.
      For $i$=$n$ to 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 one of the labels found in $w_{ij}$,
             Replace $w_i$ with $w_{ij}$, where $j$=1 to $m$.
             Update $n$=$n$+$m$-1.
         End If
      End For
    End If
```

B. Combine Separated Organization Name Words

Some organization names contain commas. Therefore, the organization names are separated into several pieces after “Section IV.A. Affiliation Separation” step. Table VIII shows some examples containing commas.

TABLE VIII
ORGANIZATION NAMES WITH COMMA SEPARATORS.

Affiliations
Institute of Optics, Information and Photonics, University Erlangen-Nuremberg, Erlangen, Germany.
Department of Traumatology, Hand and Reconstructive Surgery, Universitätsklinikum Jena
Division of Gastroenterology, Hepatology, and Nutrition, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA.
Division of Pulmonary, Critical Care Medicine, Clinical Immunology, and Allergy, Department of Internal Medicine, The David Geffen School of Medicine at UCLA, Los Angeles, CA 90095, USA.
National Heart, Lung, and Blood Institute, Epithelial Systems Biology Laboratory, National Institutes of Health, Bethesda, MD 20892, USA.
Medical Division, Japan Labor, Health, and Welfare Organization, Kawasaki-shi, Japan. hs440312@yahoo.co.jp
School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, Arizona 85287, USA.
Molecular, Cellular, and Developmental Biology Department, University of California at Santa Barbara, Santa Barbara, USA.
Department of Otology and Skull Base Surgery, Eye, Ear, Nose, and Throat Hospital, Fudan University, Shanghai, China.
Institute for Diabetes, Obesity and Metabolism, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104-5160, USA.
Office of Cellular, Tissue, and Gene Therapies, Center for Biologics Evaluation and Research, US Food and Drug Administration,
Monoclonal Antibody Research Center, Avicenna Research Institute, Academic Center for Education, Culture, and Research, Tehran, Iran.

In the case of “Institute of Optics, Information and Photonics, University Erlangen-Nuremberg, Erlangen, Germany” as shown the first row of Table VIII, the institute name (Institute of Optics, Information and Photonics) is divided into two pieces because of a comma separator. In the case of “National Heart, Lung, and Blood Institute, Bethesda, NIH, MD 20892-1603, USA”, the institute name is also divided into three pieces. We analyze all organization names and categorize them into fourteen types. Table IX shows the types that have commas in the names. In the table, OrgWord means words such as Department, School, Institute, University, Agency, etc. and w_i means any word(s).

To combine the separated names, we adopt Stanford CoreNLP tool and heuristic rules. We use two libraries in the tool. Stanford Log-linear Part-Of-Speech Tagger (POST) [14] assigns each word as noun, verb, adjective, etc. Stanford Named Entity Recognition (NER) [8] labels sequences of words in a text as person, organization, location, etc. Table X shows examples of the output. We use three tags from the POST output (second row): NNP for noun, IN for preposition/subordinating conjunction (of, for, etc.), and CC for coordinating conjunction (and, but, or, etc.). We also use three tags from NER output (third row): ORGANIZATION for organization word, LOCATION for city, state, and country, and O for other words. Table XI shows results from the two libraries for an organization name (The name is Type 1 in Table IX). The first row is the input name and the name is divided into two pieces because of a comma in the name. The second row is the POST output and the third row is the NER output. The fourth row means a word is labeled as NO if either POST output is NNP (second row) or NER output is Org (third row). The fifth row means a word is labeled as T if it is in the organization name word list (Table V) and the word is labeled as F if it is not (i.e, "Institute" is the only word in the table.). In the sixth row, NOT means a word has NO in the fourth row and T in the fifth row, and NOF means a word has NO in the fourth row and F in the fifth row.

Based on Table XI, we make a rule for the example (the first row in Table XII). If a word pattern is "NOT IN NOF^n, NOF^n CC NOF^n", combine the two pieces. Table XII shows all heuristic rules that we design to combine the split organization names. In the table, NOF^n means NOF must have appeared one or more times.

TABLE IX
ORGANIZATION NAME TYPES HAVING COMMA.

Type	Format
1	OrgWord of w_1, w_2 and w_3
2	OrgWord of w_1, w_2, and w_3
3	OrgWord of w_1, w_2, w_3 and w_4
4	OrgWord of w_1, w_2, w_3, and w_4
5	w_0 OrgWord of w_1, w_2 and w_3
6	w_0 OrgWord of w_1, w_2, and w_3
7	w_0 OrgWord of w_1, w_2, w_3 and w_4
8	w_0 OrgWord of w_1, w_2, w_3, and w_4
9	w_1, w_2 and w_3 OrgWord
10	w_1, w_2, and w_3 OrgWord
11	w_1, w_2, w_3 and w_4 OrgWord
12	w_1, w_2, w_3, and w_4 OrgWord

TABLE X
RESULTS OF POS TAGGER AND NER.

Input Affiliation	Institute of Optics, Information and Photonics, University Erlangen-Nuremberg, Erlangen, Germany.
POS Tagger (POST)	Institute/NNP of/IN Optics/NNP,/, Information/NNP and/CC Photonics/NNP ,/, University/NNP Erlangen-Nuremberg/NNP ,/, Erlangen/NNP ,/, Germany/NNP ./.
NER	Institute/ORGANIZATION of/ORGANIZATION Optics/ORGANIZATION,/O Information/ORGANIZATION and/ORGANIZATION Photonics/ORGANIZATION,/O University/ORGANIZATION Erlangen-Nuremberg/ORGANIZATION,/O Erlangen/LOCATION,/O Germany/LOCATION./O"

TABLE XI
RESULTS OF POS TAGGER AND NER (ORG MEANS ORGANIZATION).

Name	Institute	of	Optics,	Information	and	Photonics
POST	NNP	IN	NNP	NNP	CC	NNP
NER	Org	Org	Org	Org	Org	Org
NNP or Org = NO	NO		NO	NO		NO
Table V (T)	T	F	F	F	F	F
(NO or '')+T	NOT		NOF	NOF		NOF

TABLE XII
HEURISTIC RULES TO COMBINE SPLIT ORGANIZATION NAME WORDS.

Rule	Format
1	NOT IN NOF^n, NOF^n CC NOF^n
2	NOT IN NOF^n, NOF^n, CC NOF^n
3	NOT IN NOF^n, NOF^n, NOF^n CC NOF^n
4	NOT IN NOF^n, NOF^n, NOF^n, CC NOF^n
5	NOF^n NOT IN NOF^n, NOF^n CC NOF^n
6	NOF^n NOT IN NOF^n, NOF^n, CC NOF^n
7	NOF^n NOT IN NOF^n, NOF^n, NOF^n CC NOF^n
8	NOF^n NOT IN NOF^n, NOF^n, NOF^n, CC NOF^n
9	NOF^n, NOF^n CC NOF^n NOT
10	NOF^n, NOF^n, CC NOF^n NOT
11	NOF^n, NOF^n, NOF^n CC NOF^n NOT
12	NOF^n, NOF^n, NOF^n CC, NOF^n NOT

C. Possibilities of Labels Adjustment

Heuristic rules are used to adjust possibilities of City, State, Department, School, and University labels. The rules and ratios used are generated based on statistics from the training set. Table XIII shows examples of some of these rules. Rule 1 means if a word (w_i) does not explicitly suggest a City, but the previous word (w_{i-1}) indicates a possibility for University, and the next word (w_{i+1}) has a possibility for Postal Code, the word (w_i) has 80% of possibility for City.

TABLE XIII
HEURISTIC RULES FOR ADJUSTING POSSIBILITY OF A LABEL

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

D. Markov Model (MM)

Markov Model [15] is one of the most popular algorithms used for modeling time series data and used in speech recognition, gesture recognition, etc. Since labels in affiliations appear in sequence, MM is used to estimate a pattern of labels from affiliations. Equation (1) is used for MM.

$$\mathrm{P}(S_{1:T}, Y_{1:T}) = \mathrm{P}(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t) \qquad (1)$$

where Y_t is the observation at time t, S_t is the state at time t, and Y_t is independent of the states and observations at all other time indices. $X_{1:T}$ means $X_1, \ldots, X_T$ In our case, MM has seven states (S_t) (Other, Organization, City, State, Postal Code, Country, and Email) and Y_t becomes an affiliation piece of index order t. For example, "Department of History, McMaster University" is divided into two pieces. Therefore, Y_1="Department of History" and Y_2 = "McMaster University". S_t can be observed as one of the seven labels. To train MMs, we group the training set data by the label order (as shown in Table I) and train MMs for each type.

The Viterbi algorithm [16, 17] is a dynamic algorithm that computes the most probable state path through a trellis given a set of observations. Therefore, Viterbi algorithm is used to finalize the labels of each piece (word(s)) in affiliations from the MM results.

The following is the complete procedure in the proposed algorithm.

Step 1. Standardize words using Table II.
Step 2, Divide an input affiliation into several pieces using the separators in Section IV.A.
Step 3. Combine separated organization name pieces using Table XI.
Step 4. Estimate possibilities of all labels for each piece using Tables III, IV, V, VI, and VII.
Step 5. Adjust possibility of labels using Table XIII.
Step 6. Apply all trained MMs for the input affiliation and select one MM (MM_{final}) that has the highest value.
Step 7. Apply the Viterbi algorithm to MM_{final} to estimate labels of the pieces (words) in the affiliation.

V. EXPERIMENTAL RESULTS

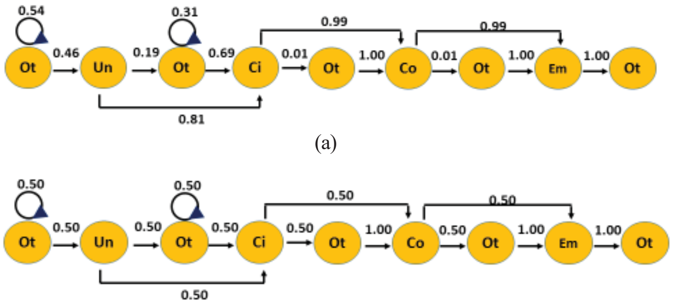
We collect 8,132 affiliations from MEDLINE citations published from 1985 to 2014. From these, 4,446 are associated for training and 3,686 for testing. In MEDLINE, each citation has its unique ID called PubMedID (PMID). We collect all affiliations in PMIDs ranging from 23,000,000 to 23,005,000 for training and PMIDs ranging from 23,005,001 to 23,010,000 for testing. Since some PMIDs do have citations and some citations do not have affiliations, the training and testing sets have different sizes.

To optimize the number of MMs and number of training data for each MM, we first remove “Other” that occurs between other labels in the training data, group the training data by the order of labels, and train MMs for each group. For example, “Other, University, Other, City, State, Country” is assigned to “Other, University, City, State, Country” group for training.

We have 61 MMs from the training set. Table XIV shows some of the trained MMs. Some MMs contain a reasonable size of training data. Twenty-six MMs have more than ten affiliations (the first five models in the table). However, thirty-five MMs have less than ten affiliations. Among them, thirteen MMs have only one affiliation. The bottom four MMs in the table show the examples. The second row has 597 training data and the corresponding diagram of the MM is shown in Fig. 1. There are two diagrams in the figure. Fig. 1(a) shows the MM from the training data (MM Trained) and Fig. 1(b) shows the MM modified from the MM 1(a) (MM Equal weight). The transition workflows from one label to other labels are the same in the two MMs. The difference is that Fig. 1(b) has equal transition weights (=1/k) when one label can move to k different labels. Since MM is a probabilistic module, MM (MM Trained) is trained in favor of most frequent cases. Therefore, the MM does not perform well for less frequent input cases. To resolve this issue, I design a new MM (MM Equal weight).

TABLE XIV
MMS TRAINED USING THE TRAINING SET

MM	Number of affiliations used
Other,University,City,Country	724
Other, University, City, Country, Email	597
Other, University, City, Postal Code, Country	505
Other, University, City, State, Country	314
Other, University, Country	116
Other, University, Postal Code, City, Country	5
Other, University, Postal Code, City	4
Other, University, Postal Code, Email	2
Other, University, State	1



(a)

(b)

Figure 1. (a) MM (MM Trained) in the third row in Table XIV. (b) MM (MM Equal weight) from the MM (a). The meaning of the abbreviations (Ot, Un, Ct, etc.) are the same as the ones in Table I.

Table XV shows the test results. We consider errors when one of the labels in an affiliation is mislabeled by MMs. The MMs Trained shows 95.90% accuracy and MMs Equal weight shows 85.59% accuracy. The MMs Trained shows better performance than the MMs Equal weight. Since the MMs Equal weight does not consider statistics, the MMs does not have good performance overall. However, the MMs Equal weight has the same weight for all possible paths. Therefore, MMs Equal weight will have better performance for affiliations that have less frequent label orders. We compare our results with Yu et al. [4]. They have 94.0% accuracy for estimating Country and 87.0% accuracy for Institution from affiliations. Since we estimate seven labels and consider as an error when one of labels is mislabeled in an affiliation, our method shows relatively good performance.

We analyze the 151 errors from MMs Trained as shown in the Table XV. The first issue is the confusion of University level organization names. In the first row, the MMs labeled “Pomerado Hospital” as University instead of “Palomar Pomerado Health”. Since “Palomar Pomerado Health” does not contain any clue words related to organization name and University level organization name does not always show at the end, the MMs create the error. The second row shows an affiliation separation error. The proposed algorithm cannot process affiliations correctly when there is no proper separator in affiliations. The third row shows an affiliation written in Polish. There are no separators between department and university, and between postal code and city name. There is no country name there. All these issues make it difficult for the algorithm to separate the words. The fourth row shows the proposed "Combine Separated Organization Name Words (Section IV.B)" algorithm error. "Department of Nutrition and Food Sciences, Physiology and Toxicology" should be combined as a piece. However, it is divided into two pieces ("Department of Nutrition and Food Sciences" and "Physiology and Toxicology") because of two "and" words. There is no heuristic rule for this case in Table XII. The fifth row shows two affiliations in the text. Since existing MMs are trained for processing an affiliation, the MMs cannot handle texts with multiple affiliations. The sixth row shows an affiliation of a private company. Since the organization name does not contain any clue words related to organization names, the algorithm created a labeling error. This problem can be resolved by collecting University level names.

TABLE XV
PERFORMANCE OF THE PROPOSED MMS

MM Mode	MMs Trained	MMs Modified
Total Affiliations	3,686	3,686
True	3,535	3,273
False	151	551
Accuracy	95.90%	85.59%

TABLE XVI
ERROR ANALYSIS

Error Analysis	Affiliation Example
Confusion of University level organization name	Pomerado Hospital, Palomar Pomerado Health, 15615 Pomerado Road, Poway, CA 92064, USA. david.tam@pph.org
No separator	Department of Educational Psychology in the College of Education at the University of Washington.
Foreign language with no separators	Klinika Chirurgii Naczyniowej, Ogólnej i Angiologii Pomorskiego Uniwersytetu Medycznego w Szczecinie al. Powstanców Wlkp. 72, 70-111 Szczecin.
Combining algorithm error	Department of Nutrition and Food Sciences, Physiology and Toxicology, University of Navarra, Pamplona, Spain
Multiple affiliations	Department of Physics, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal and Centro de Física Nuclear, Universidade de Lisboa, 1649-003 Lisbon, Portugal.
Unknown organization name	Bolle Safety/Bolle Tactical, Bushnell Outdoor Products. pkroesch@bushnell.com

VI. CONCLUSIONS

This paper proposes an automatic labeling algorithm to classify labels from affiliations in biomedical journal articles using Stanford CoreNLP tool, MM, Viterbi algorithm, statistics, and heuristic rules. This is a necessary step to extract statistics on number of publications by geography or organization. We collect seven word list tables to estimate the probabilities of seven different labels for each word in the affiliations. We use 4,446 affiliations for a training set and 3,686 affiliations for a testing set collected from MEDLINE. The proposed modules have 95.90% and 85.59% accuracies from MMs Trained and MMs Equal weight, respectively.

As a future work, we plan to collect more organization names and corresponding labels (City, State, Country, etc.) to improve labeling accuracy and to estimate important labels missing in affiliations. In addition, we plan to collect more organization name related (foreign language) words for Table V. We also plan to improve algorithms for separating affiliation texts and combining separated affiliation pieces to improve classification accuracy.

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] Kim J, Lobuglio PS, Thoma GR, "Visualization of Statistics from MEDLINE", 2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS 2016), Dublin and Belfast, Ireland, pp. 290-291, June, 2016.

[3] 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.

[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] Torvik VI, "MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide", D-Lib® Magazine, Vol 21, No. 11/12, 2015.

[6] Bhargava R, D'Ignazio C. 2014. CLIFF Mediameter. MIT Center for Civic Media. (http://cliff.mediameter.org)

[7] 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.

[8] http://nlp.stanford.edu/software/CRF-NER.shtml.

[9] Kim J, Thoma GR. "Named Entity Recognition in Affiliations of Biomedical Articles Using Statistics and HMM Classifiers", The 2016 International Conference on Data Mining (DMIN2016), Las Vegas, USA, pp. 236-241, July, 2016

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

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

[12] http://regexlib.com.

[13] http://stanfordnlp.github.io/CoreNLP

[14] http://nlp.stanford.edu/software/tagger.html.

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

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

[17] Forney, G. D., "The Viterbi Algorithm", Proceeding of the IEEE, vol. 61, 268-278, 1973.