MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-910

A Semi-Automatic Framework to Identify Abnormal States in EHR Narratives

Xiaojun Ma[a], Takeshi Imai[a], Emiko Shinohara[b], Ryota Sakurai[b], Kouji Kozaki[c], Kazuhiko Ohe[a,b]

[a] Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
[b] The University of Tokyo Hospital, Tokyo, Japan
[c] The Institute of Scientific and Industrial Research, Osaka University, Osaka, Japan

Abstract

Disease ontology, defined as a causal chain of abnormal states, is believed to be a valuable knowledge base in medical information systems. Automatic mapping between electronic health records (EHR) and disease ontology is indispensable for applying disease ontology in real clinical settings. Based on an analysis of ontologies of 148 chronic diseases, approximately 41% of abnormal states require information extraction from clinical narratives. This paper presents a semi-automatic framework to identify abnormal states in clinical narratives. This framework aims to effectively build mapping modules between EHR and disease ontology. We show that the proposed method is effective in data mapping for 18%-33% of the abnormal states in the ontologies of chronic diseases. Moreover, we analyze the abnormal states for which our method is invalid in extracting information from clinical narratives.

Keywords:

Ontology; Machine learning; Natural language processing

Introduction

In order to understand diseases, one must adequately capture abnormal states in diseases. With support from Japan's Ministry of Health, Labour and Welfare, we have been involved in developing a disease ontology [1,2], wherein a disease is captured as a causal chain of abnormal states. So far, medical experts have described the causal chains of approximately 6,000 diseases across 13 medical departments. The disease ontology provides domain-specific knowledge, answering questions such as "what abnormal states cause a disease?" and "how might a disease progress, and what symptoms may appear?". We believe that disease ontology is a valuable knowledge base in medical information systems such as deep phenotyping in support of precision medicine. For the practical use of disease ontology in clinical applications, automated mapping between electronic health records (EHR) and disease ontology is indispensable. An investigation has been conducted to specify clinical data sources where information can be extracted to identify abnormal states. The investigation is based on an analysis of 4,718 (643 distinct) abnormal states of 148 chronic diseases. We categorize data sources in EHR as clinical notes, exam reports, laboratory data, treatment orderings, and demographics, which each require different mapping techniques. From the investigation results, approximately 41% of the abnormal states are supposed to be identified in unstructured clinical notes or exam reports. Therefore, our study first explores techniques of mapping clinical narratives to abnormal states, which involves machine learning and natural language processing (NLP).

In previous studies, Informatics for Integrating Biology and the Bedside (i2b2) has organized NLP challenges including identifying patient smoking [3] and obesity [4] status from English-language discharge summaries. A common method exploited by the participants can be summarized in two steps: (1) retrieve relevant sentences or passages via keyword or rule-based search, (2) build sentence-level or passage-level classifiers to determine status, which we refer to here as polarity classification. In this paper, we present a general framework, extended from the conventional method, for extracting information from clinical narratives to identify abnormal states in disease ontology. The difference between our study and the i2b2's challenges lays in the following aspects: (1) all available clinical notes in EHR are used in our study including discharge summaries, progress notes, nursery notes, which incorporate a wide variety of narratives; (2) the clinical narratives in our study are written in Japanese, which requires Japanese language processing such as morphological analysis. The improvements we have made to the method include the following: (1) adopting a semi-automatic keyword expansion for candidate sentence retrieval, using both knowledge-driven and data-driven approaches; (2) applying Synthetic Minority Over-sampling Technique (SMOTE) [5] to solve class imbalance problem in polarity classification; and (3) introducing a mechanism of using existing labeled data to automatically label unlabeled data, which reduces the labeling costs of machine learning for polarity classification.

Our study focuses on chronic diseases across the life span including diabetes, hypertension, dyslipidemia, chronic kidney disease, and ischemic heart disease. The objective of our study is to propose a general framework to efficiently build mapping modules between clinical narratives and abnormal states in ontologies of chronic diseases. In this paper, we discuss the technical requirements to identify abnormal states in clinical narratives, and we validate the applicability of the proposed framework using EHR data from the University of Tokyo Hospital. Experiments demonstrate that the proposed method is effective in EHR data mapping for 18%-33% of the abnormal states in the ontologies of chronic diseases.

The rest of the paper is organized as follows. The following section introduces the results of the investigation into abnormal states. The Methods section describes the proposed framework. The Experiment Section describes experiments designed to evaluate the performance. In the Discussion section, we discuss the proposed method and those abnormal states whose mapping modules cannot be realized by our framework. The final section presents our conclusions.

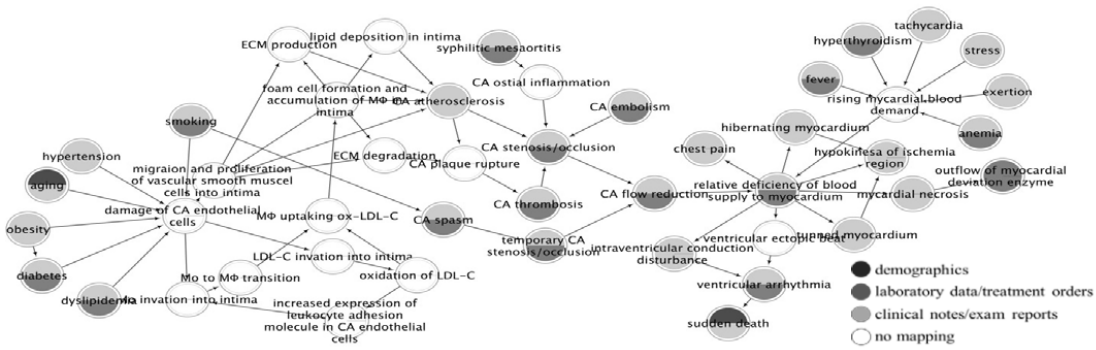

Figure 1 – Disease ontology of angina pectoris

Investigation of Abnormal States

Based on an analysis of 4,718 (643 distinct) abnormal states of 148 chronic diseases by medical experts, we found that 384 abnormal states (59%) are potentially able to be mapped to combinations of clinical information, such as clinical notes, exam reports, laboratory data, treatment orders, and demographics (see Table 1). Figure 1 gives an example of disease ontology for angina pectoris with descriptions of categories of data sources. Laboratory data, treatment orders, and demographics are structured data, which are easy to access. Clinical notes and exam reports are in unstructured narrative form, where approximately 41% of the abnormal states are supposed to be identified.

We further analyze the abnormal states that require information extraction from narratives. We classify them into overlapping categories of "easy" and "difficult". The "easy" abnormal states are described by word-level expressions in EHR and have concrete clinical meanings, which can be identified by terms like synonyms, hypernyms or hyponyms. The "difficult" abnormal states are those described at sentence level, or expressed in abstract terms, or that require numerical extraction from text. The distributions of "easy", "easy/difficult", and "difficult" abnormal states are shown in Table 2. We believe that the general framework proposed in this paper is applicable in mapping EHR narratives to "easy" ones, which accounts for 18%-33% of all the abnormal states in the ontologies of chronic diseases.

Table 1 – Data sources for identifying abnormal states

Data sources categories	No. of abnormal states	Rates (%)
Possible mapping from EHR	382	59.4
Clinical notes/ Exam reports	263	40.9
Laboratory data	131	20.4
Treatment orders	43	6.7
Demographics	2	0.3
No evidence in EHR	261	40.6
All	643	-

Table 2 – Number of "easy", "easy/difficult", and "difficulty" abnormal states

Easy	Easy or Difficult	Difficult
115	100	48

Methods

The workflow of the proposed framework for building mapping modules between clinical narratives and abnormal states is shown in Figure 2.

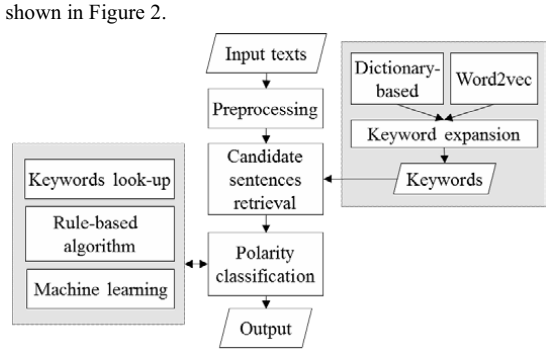

Figure 2 – Workflow of the proposed framework

Preprocessing

Input texts are normalized, which involves full and half width conversion, upper/lower case mapping, and orthographical variants correction. The normalized texts are split into paragraphs based on line break and then paragraphs are decomposed into sentences based on Japanese punctuation characters.

Candidate Sentence Retrieval

Sentences related to an abnormal state are retrieved by keyword search with duplicates removed. We first make a list of keywords manually for an abnormal state and then implement keyword expansion to generate a set of keywords for sentence retrieval. If a keyword has known irrelevant multiple concepts, search rules are made to exclude irrelevant sentences from search results. For keyword expansion, two kinds of approaches are taken. One is dictionary-based, which implements synonym keyword expansion based on concept unique identifiers (CUI) of the Unified Medical Language System (UMLS: Japanese MeSH, Japanese MedDRA) [6]. The other is to train word2vec [7] to obtain similar words in learned vector space. We train two word2vec models using two corpora. One model is trained on a corpus consisting of Japanese medical tutorials including one Japanese medical textbook, Merck Manual professional [8] and home editions [9], and Wikipedia pages related to medicine.

The other model is trained on real EHR narratives using a dataset described as DS2 in the Experiments section. Japanese morphological analysis (JMA) is conducted to generate word

tokens beforehand, among which nouns are used to train word2vec models. The JMA is based on a customized medical dictionary.

Polarity Classification

For each retrieved sentence, affirmation or negation (polarity) of an abnormal state is determined. It is a classification problem, where three methods are applied to build classifiers.

Keyword Look-up (KL)

With respect to some keywords, their presence can directly indicate polarity. For example, the existence of "bi=" ("bi" is an abbreviation of "brinkman index") implies that the patient is a smoker.

Rule-based Algorithm (RA)

Regular expressions are created to determine the polarity of sentences retrieved by some keywords, especially English keywords such as "smoker", "smoking", and "obesity". In the case of abnormal states of disease names such as "diabetes", and "dyslipidemia", it is rare to see their negations and the negative expressions are quite similar across diseases. Therefore, limited negative expressions can be collected for polarity classification and shared by multiple disease names.

Machine Learning (ML)

Sentences related to symptoms and those related to lifestyle issues retrieved by Japanese keywords show great variety in the contexts in which they appear. In these cases, machine learning is applied.

We use linear support vector machine (SVM) as a base algorithm and bag-of-words (BoW) as features. To construct a BoW feature for each sentence, the JMA is conducted to tokenize each sentence into words. The window size of words used to construct features is adjusted for the best performance.

For an abnormal state, affirmative sentences are often much more common than negative ones. For those data with severe class imbalance problem, SMOTE is applied.

Input
$\{s_i\} \in S$: set of labeled data in source domain
$\{t_j\} \in T$: set of unlabeled data in target domain
Output
$T^l \in T$: set of labeled data in target domain
Procedure
1. Train a KNN model on *S*.
2. Use the model to annotate each $\{t_j\} \in T$ to get h_j: $t_j \rightarrow$ label by majority voting of the most *k* closest s_i with probability. For $j = 1,\ldots,|T|$ do:
  If $prob(h_j) = 1$ then
    Set label of t_j
    with h_j
    $T^l = T^l \cup t_c$

Figure 3 – Mechanism for RLC

Labeling data is expensive. We try to reuse existing labeled data of an abnormal state (source domain) to reduce the workload of labeling unlabeled data for a new abnormal state (target domain). We establish a mechanism for reducing manual labeling, denoted as RLC, as described in Figure 3. A k-nearest neighbors (KNN) model trained on labeled source-domain data is applied to annotate unlabeled target-domain data. The labels outputted with high probability are assigned to the data in the target domain automatically, while the rest of the unlabeled data remain for manual labeling

Experiments

Data Description

The data for this study is based on EHR data from the University of Tokyo Hospital, involving a great variety of types of clinical notes, including discharge summaries, progress notes, nursery notes. Exam reports are usually transcribed in clinical notes, so the data we used includes texts of both clinical notes and exam reports.

Dataset 1 (DS1)

We collected all clinical notes generated in 2015 at the University of Tokyo Hospital, which is denoted as DS1. DS1 is composed of approximately 170,000 clinical notes, which we used to evaluate our method.

Dataset 2 (DS2)

In order to train word2vec model for keyword expansion, we collected a dataset denoted as DS2. DS2 consists of approximately 60,000 clinical notes generated from 2010 to 2015 at the University of Tokyo Hospital, which contain at least one of the following disease names: chronic kidney disease, is- chemic heart disease, and arteriosclerosis.

Experimental Settings and Evaluation Methods

Two things we concern are: first, "how precise is identification of an abnormal state", and second, "how many patients with a given abnormal state are recalled".

Evaluation of polarity classification answers the first question. We constructed a KL or RA by viewing approximately 20 candidate sentences. For each KL or RA classifier, if the total number of candidate sentences in DS1 is much higher than that used for classifier construction, we sampled approximately 100 sentences to create test data for evaluation. To construct a training set for a ML, we randomly sampled 250 sentences, if available, from candidate sentences retrieved from DS1. For those abnormals for which there were not enough candidate sentences, all of the available sentences were used. The training sets were manually labeled and five-fold cross validation was conducted to evaluate the performance of ML classifiers.

A direct answer to the second question is achieved by computing patients' recall for every abnormal state. This requires annotation by medical experts, which is extremely expensive in terms of both time and financial cost. In our proposed method, recall depends on the effectiveness of keyword expansion. Therefore, at this stage, we present increased ratios of retrieved clinical notes by keyword expansion instead.

Experimental Results

We applied our proposed framework to 10 abnormal states, which are summarized in Table 3.

Effectiveness of Keyword Expansion

The keyword expansion grew the keyword lists for candidate sentence retrieval. Table 3 shows the increased percentages of retrieved notes from DS1 compared to those retrieved before keyword expansion (see column "Eva on KE").

Performance of Polarity Classification

Table 3 gives the methods of polarity classification for each abnormal state and their performance measured in marco-averaged precision (Avg. P), recall (Avg. R), and F1 score (Avg. F1), as well as classification accuracy (CA).

For ML, we investigated the effect of dataset size on macro-averaged F1 scores of five-fold cross validation (see Figure 4). We found that for most abnormal states with dataset size above 100-250, F1 increases more slowly and standard deviation of

F1 converges. The observation is basically consistent with a previous study [10] of the i2b2 Smoking Challenge, which states that accuracy increases slowly when the training set size is over 200 using hold-out validation. Therefore, in this study, we used 250 labeled sentences for each ML (except for the abnormal state of pulmonary congestion which has only 142 candidate sentences available in DS1). The performance shown in Table 3 for ML is based on fully manually annotated data without introducing RLC.

Table 3 – Effectiveness of keyword expansion and performance of polarity classification

Abnormal States	Eva on KE	Methods	Polarity Classification Avg. P	Avg. R	Avg. F1	CA
diabetes	52.0%	RA	-	-	-	0.98
dyslipidemia	68.6%	RA	-	-	-	0.99
obesity	37.9%	KL, RA	1.00	1.00	1.00	1.00
smoking	1.1%	KL,RA,ML	0.94	0.95	0.94	0.95
anura	0.0%	KL	-	-	-	-
chest pain	52.7%	ML	0.93	0.95	0.94	0.95
fever	36.7%	RA, ML	0.93	0.97	0.95	0.96
overhydration	41.4%	ML	0.89	0.91	0.90	0.91
pulmonary congestion	29.3%	ML	0.90	0.92	0.91	0.92
tachycardia	186.2%	ML	0.95	0.96	0.96	0.96

* Negation of a disease name is extremely rare, so only CA is given.

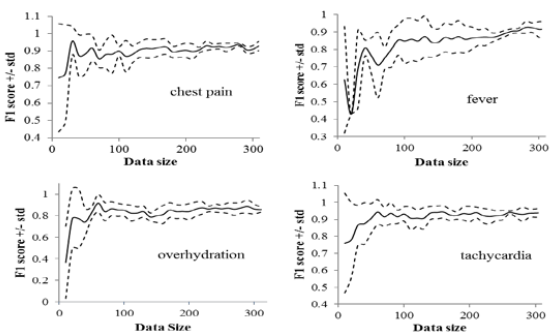

Figure 4 – Effect of dataset size on F1 scores

Availability of RLC

In the experiments on polarity classification using the fully manually annotated data, ML models for chest pain, fever, overhydration, pulmonary congestion and tachycardia showed the best performance with the same word count window in feature construction, which uses eight tokens after the keywords. All these abnormal states are symptoms. We assume that their candidate sentences may have similar contexts. A new experiment was designed to verify RLC based on these five abnormal states of symptoms.

In order to measure similarity between feature spaces of retrieved sentences containing the five symptoms, we computed the cosine similarity between word count vectors of the 250 annotated sentences for every two symptoms, which is shown in Table 4. Table 5 describes the performance (F1 scores) of the ML models trained on data of one symptom to predict polarity for the other four symptoms. The Pearson correlation coefficient between Table 4 and Table 5 is 0.422, indicating moderate positive correlation.

We applied RLC to the five abnormal states of symptoms, using labeled data of one symptom (source domain) to reduce the labeling costs of the other four (target domain). The rates of reduced labeling costs, that is, the proportions of automatically labeled data in the target domains, are shown in Table 6, ranging from 1.6% to 69.7%, averagely 32.4%. Table 7 gives the macro-averaged F1 scores of five-fold cross validation of polarity classification using the RLC outputs, which is similar to the performance using the fully manually annotated data as shown in Table 3.

Table 4 – Cosine similarity between word count vectors for every two symptoms

	chest pain	fever	over-hydration	pulmonary congestion	tachycardia
chest pain	-	0.89	0.76	0.82	0.91
fever	-	-	0.79	0.84	0.85
overhydration	-	-	-	0.79	0.83
pulmonary congestion	-	-	-	-	0.82
tachycardia	-	-	-	-	

Table 5 – Performance of polarity classifiers trained on data of one symptom to predict others

Train \ Test	chest pain	fever	over-hydration	pulmonary congestion	tachycardia
chest pain	-	0.87	0.73	0.88	0.9
fever	0.92	-	0.77	0.86	0.93
overhydration	0.9	0.88	-	0.9	0.89
pulmonary congestion	0.92	0.87	0.74	-	0.91
tachycardia	0.9	0.88	0.73	0.89	-

Table 6 – Percentage of reduced labeling costs

Source \ Target	chest pain	fever	over-hydration	pulmonary congestion	tachycardia
chest pain	-	33.6	16	19.7	42
fever	43.6	-	40	49.3	50
overhydration	13.2	14	-	12	21.2
pulmonary congestion	11.6	10	1.6	-	14
tachycardia	61.2	63.6	60.8	69.7	-

Table 7 – Performance of polarity classification using RLC

Source \ Target	chest pain	fever	over-hydration	pulmonary congestion	tachycardia
chest pain	-	0.96	0.88	0.90	0.96
fever	0.94	-	0.86	0.92	0.97
overhydration	0.93	0.96	-	0.91	0.96
pulmonary congestion	0.93	0.96	0.86	-	0.96
tachycardia	0.95	0.96	0.89	0.95	-

Discussion

Applicability of the Proposed Framework

Our proposed framework is useful in mapping clinical narratives to abnormal states that are described by word-level expressions and have concrete clinical meanings. Keyword expansion and polarity classification are two key processes in our method.

Table 3 shows that the keyword expansion increases the number of retrieved clinical notes by 51% on average, which demonstrates the effectiveness of keyword expansion in improving recall. We also found that word2vec is able to find candidate keywords that the dictionary-based method cannot obtain, such as orthographical variants, obsolete terms but still in use, and informal expressions used locally in each hospital. However, keyword candidates outputted by word2vec show low precision, which requires manual selection. In future work, refining word representation learned from corpora by using semantic lexicons [11] may be a promising approach to outputting keyword candidates more efficiently.

As for the polarity classification, the proposed framework solves two problems in building ML classifiers: class imbalance and labeling costs. We applied SMOTE to solve the former problem, improving the macro-averaged F1 scores of cross validation by 5% to 10%. The F1 scores with SMOTE applied

are shown in Table 3. For the latter problem, we introduced RLC. Figure 4 shows that the size of the training data set necessary to build a high-performance ML classifier is above 100-200. Labeling training data for every abnormal state is time-consuming. RLC reuses existing labeled data to automatically label those unlabeled data that are similar to the existing labeled ones, thus reducing labeling costs. From Table 4 and Table 5, when the training and the test data come from different domains, moderate positive correlation exists between the cosine similarity of their feature spaces and the performance of the models. It is reasonable to consider using data in source domain to annotate data in target domain if they have similar feature spaces. Based on the experiments on the five abnormal states of symptoms, Table 6 and Table 7 demonstrate that RLC can reduce the labeling costs by 32.4% on average, achieving similar performance as using the fully manually annotated data. The rate of reduced labeling costs possibly depends on the dataset size of the source domain, and the similarity between feature spaces of the source and target domain. RLC is a simple and effective way of reducing labeling costs in our proposed framework.

Limitations of the Proposed Framework

Our proposed framework is useful in mapping clinical narratives. The proposed framework is invalid in identifying some abnormal states whose information is supposed to be contained in clinical narratives. We summarize these kinds of abnormal states as follows.

Some abnormal states are expressed at sentence level in EHR like “feel pain in chest”, and “the spleen is enlarged”. These kinds of expressions are missed in candidate sentence retrieval which is based on keyword search in our proposed method.

Abnormal states like “stress” and “systematic inflammation” have abstract clinical meanings, which causes difficulty in collecting keywords for retrieving candidate sentences.

Some abnormal states such as “exertion” usually appear with other states such as “chest pain upon exertion”, and “difficult respiration upon exertion”. Thus, retrieved sentences do not focus on “exertion” but respond to multiple abnormal states. Therefore, mapping module should be established between EHR and a cluster of abnormal states in disease ontology.

Extraction of numerical information such as body weight, body mass index (BMI), and blood pressure is necessary for identifying some abnormal states like “obesity” and “hypertension”. Numerical extraction from narratives requires more sophisticated NLP techniques.

Future Directions

Our future work will include the following: (1) building modules of mapping other data sources such as laboratory data and treatment orders to disease ontology; (2) developing sophisticated NLP techniques to identify the abnormal states in narratives that cannot be achieved by the proposed method; (3) exploiting structures in disease ontology to infer those abnormal states that have no trace of evidence in EHR.

Conclusions

In this paper, we have proposed a general framework to identify abnormal states in EHR narratives, which is applicable in data mapping for 18%-33% of the abnormal states in the ontologies of chronic diseases. The proposed method exhibits improvement over the conventional method by applying keyword expansion based on UMLS and word2vec, and by introducing SMOTE and RLC in ML for polarity classification.

Acknowledgements

This research is partially supported by the ICT infrastructure establishment for clinical and medical research from Japan Agency for Medical Research and development, AMED, and Health and Labour Sciences Research Grants from the Ministry of Health, Labour and Welfare, Japan.

References

[1] K. Kozaki, R. Mizoguchi, T. Imai, and K. Ohe, Identity tracking of a disease as a causal chain, *Proc. ICBO* (2012), 131-136.

[2] K. Kozaki, R. Mizoguchi, T. Imai, and K. Ohe, A consid- eration on identity of diseases. *Proc. InterOntology* (2012), 75-80.

[3] O. Uzuner, I. Goldsten, Y. Luo, and I. Kohane, Identifying patient smoking status from medical discharge records, *JAMIA* **15** (2008), 14-24.

[4] O. Uzuner, Recognizing obesity and comorbidities in sparse data, *JAMIA* **16** (2009), 561-570.

[5] N.V. Chawala, K.W. Bowyer, L.O. Hall, and W.P Keg- elmeyer, SMOTE: Synthetic minority over-sampling tech- nique, *JAIR* **16** (2002), 321-357.

[6] Unified medical language system (UMLS). Available at: https://www.nlm.nih.gov/research/umls/. [accessed Dec. 2016]

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, *NIPS* (2013), 3111-3119.

[8] Merck Manual Professional Edition in Japanese. Available at: http://merckmanual.jp/mmpej/index.html. [accessed Dec. 2016]

[9] Merck Manual Home Edition in Japanese. Available at: http://merckmanuals.jp/home/index.html. [accessed Dec. 2016]

[10] C. Clark, K. Good, L. Jezierny, M. Macpherson, B. Wil- son, and U. Chajewska, Identifying smokers with a medi- cal extraction system, *JAMIA* **15** (2007), 36-39.

[11] M. Faruqui, J. Dodge, S.K. Jauhar, C. Dyer, E. Hovy, and N.A. Smith, Retrofitting word vectors to semantic lexi- cons, *Proc. NAACL-HLT* (2015), 1606-1615.

Address for correspondence

Xiaojun Ma, Department of Biomedical Informatics, Graduate School of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655 Japan.

Email:xiaojun_ma@m.u-tokyo.ac.jp