Hayato Ohwada
Kenichi Yoshida (Eds.)

LNAI 9806

Knowledge Management and Acquisition for Intelligent Systems

14th Pacific Rim Knowledge Acquisition Workshop, PKAW 2016
Phuket, Thailand, August 22–23, 2016
Proceedings

Springer

Lecture Notes in Artificial Intelligence 9806

Subseries of Lecture Notes in Computer Science

More information about this series at http://www.springer.com/series/1244

Hayato Ohwada · Kenichi Yoshida (Eds.)

Knowledge Management and Acquisition for Intelligent Systems

14th Pacific Rim
Knowledge Acquisition Workshop, PKAW 2016
Phuket, Thailand, August 22–23, 2016
Proceedings

Springer

Editors
Hayato Ohwada
Tokyo University of Science
Noda, Chiba
Japan

Kenichi Yoshida
University of Tsukuba
Bunkyo, Tokyo
Japan

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Artificial Intelligence
ISBN 978-3-319-42705-8 ISBN 978-3-319-42706-5 (eBook)
DOI 10.1007/978-3-319-42706-5

Library of Congress Control Number: 2016944819

LNCS Sublibrary: SL7 – Artificial Intelligence

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

This volume contains the papers presented at PKAW2016: The 14th International Workshop on Knowledge Management and Acquisition for Intelligent Systems, held during August 22–23, 2016 in Phuket, Thailand, in conjunction with the 14th Pacific Rim International Conference on Artificial Intelligence (PRICAI 2016).

In recent years, unprecedented data, called big data, have become available and knowledge acquisition and learning from big data are increasing in importance. Various types of knowledge can be acquired not only from human experts but also from diverse data. Simultaneous acquisition from both data and human experts increases its importance. Multidisciplinary research including knowledge engineering, machine learning, natural language processing, human–computer interaction, and artificial intelligence is required. We invited authors to submit papers on all aspects of these area. Another important and related area is applications. Not only in the engineering field but also in the social science field (e.g., economics, social networks, and sociology), recent progress in knowledge acquisition and data engineering techniques is leading to interesting applications.

We invited submissions that present applications tested and deployed in real-life settings. These papers should address lessons learned from application development and deployment. As a result, a total of 61 papers were considered. Each paper was reviewed by at least two reviewers, of which 28 % were accepted as regular papers and 8 % as short papers. The papers were revised according to the reviewers' comments. Thus, this volume includes 16 regular papers and five short papers. We hope that these selected papers and the discussion during the workshop lead to new contributions in this research area.

The workshop co-chairs would like to thank all those who contributed to PKAW 2016, including the PKAW Program Committee and other reviewers for their support and timely review of papers and the PRICAI Organizing Committee for handling all of the administrative and local matters. Thanks to EasyChair for streamlining the whole process of producing this volume. Particular thanks to those who submitted papers, presented, and attended the workshop. We hope to see you again in 2018.

August 2016

Hayato Ohwada
Kenichi Yoshida

Organization

Honorary Chairs

Paul Compton	University of New South Wales, Australia
Hiroshi Motoda	Osaka University and AFOSR/AOARD, Japan

Workshop Co-chairs

Hayato Ohwada	Tokyo University of Science, Japan
Kenichi Yoshida	University of Tsukuba, Japan

Advisory Committee

Byeong-Ho Kang	School of Computing and Information Systems, University of Tasmania, Australia
Deborah Richards	Macquarie University, Australia

Program Committee

Nathalie Aussenac-Gilles	IRIT CNRS, France
Quan Bai	Auckland University of Technology, New Zealand
Ghassan Beydoun	University of Wollongong, Australia
Ivan Bindoff	University of Tasmania, Australia
Xiongcai Cai	University of New South Wales, Australia
Aldo Gangemi	Université Paris 13 and CNR-ISTC, France
Udo Hahn	Jena University, Germany
Nobuhiro Inuzuka	Nagoya Institute of Technology, Japan
Toshihiro Kamishima	National Institute of Advanced Industrial Science and Technology, Japan
Mihye Kim	Catholic University of Daegu, South Korea
Yang Sok Kim	University of Tasmania, Australia
Masahiro Kimura	Ryukoku University, Japan
Alfred Krzywicki	University of New South Wales, Australia
Setsuya Kurahashi	University of Tsukuba, Japan
Maria Lee	Shih Chien University
Kyongho Min	University of New South Wales, Australia
Toshiro Minami	Kyushu Institute of Information Sciences and Kyushu University Library, Japan
Luke Mirowski	University of Tasmania, Australia
James Montgomery	University of Tasmania, Australia
Tsuyoshi Murata	Tokyo Institute of Technology, Japan

Kouzou Ohara	Aoyama Gakuin University, Japan
Tomonobu Ozaki	Nihon University, Japan
Son Bao Pham	College of Technology, VNU, Vietnam
Alun Preece	Cardiff University, UK
Ulrich Reimer	University of Applied Sciences St. Gallen, Switzerland
Kazumi Saito	University of Shizuoka, Japan
Derek Sleeman	University of Aberdeen, UK
Vojtěch Svátek	University of Economics, Prague, Czech Republic
Takao Terano	Tokyo Institute of Technology, Japan
Shuxiang Xu	University of Tasmania, Australia
Tetsuya Yoshida	Nara Women's University, Japan

Contents

Knowledge Acquisition and Machine Learning

Abbreviation Identification in Clinical Notes with Level-wise Feature Engineering and Supervised Learning

Thi Ngoc Chau Vo[1(✉)], Tru Hoang Cao[1], and Tu Bao Ho[2,3]

[1] University of Technology, Vietnam National University, Ho Chi Minh City, Vietnam
{Chauvtn, tru}@cse.hcmut.edu.vn
[2] Japan Advanced Institute of Science and Technology, Nomi, Japan
bao@jaist.ac.jp
[3] John von Neumann Institute, Vietnam National University, Ho Chi Minh City, Vietnam

Abstract. Nowadays, electronic medical records get more popular and significant in medical, biomedical, and healthcare research activities. Their popularity and significance lead to a growing need for sharing and utilizing them from the outside. However, explicit noises in the shared records might hinder users in their efforts to understand and consume the records. One kind of explicit noises that has a strong impact on the readability of the records is a set of abbreviations written in free text in the records because of writing-time saving and record simplification. Therefore, automatically identifying abbreviations and replacing them with their correct long forms are necessary for enhancing their readability and further their sharability. In this paper, our work concentrates on abbreviation identification to lay the foundations for de-noising clinical text with abbreviation resolution. Our proposed solution to abbreviation identification is general, practical, simple but effective with level-wise feature engineering and a supervised learning mechanism. We do level-wise feature engineering to characterize each token that is either an abbreviation or a non-abbreviation at the token, sentence, and note levels to formulate a comprehensive vector representation in a vector space. After that, many open options can be made to build an abbreviation identifier in a supervised learning mechanism and the resulting identifier can be used for automatic abbreviation identification in clinical text of the electronic medical records. Experimental results on various real clinical note types have confirmed the effectiveness of our solution with high accuracy, precision, recall, and F-measure for abbreviation identification.

Keywords: Electronic medical record · Clinical note · Abbreviation identification · Level-wise feature engineering · Supervised learning · Word embedding

1 Introduction

In recent years, there has been a growing need for sharing and utilizing electronic medical records from the outside in medical, biomedical, and healthcare research activities. Free text in clinical notes of these electronic medical records often contains

H. Ohwada and K. Yoshida (Eds.): PKAW 2016, LNAI 9806, pp. 3–17, 2016.
DOI: 10.1007/978-3-319-42706-5_1

explicit noises such as spelling errors, variants of terms (acronyms, abbreviations, synonyms …), unfinished sentences, etc. [8]. Such explicit noises in the shared records might hinder users in their efforts to understand and consume the records. One kind of explicit noises that has a strong impact on the readability of the records is a set of abbreviations written in free text in the records because of writing-time saving and record simplification. Those abbreviations might result in misinterpretation and confusion of the content in the electronic medical records as mentioned in [4]. So, automatically identifying abbreviations and replacing them with their correct long forms are significant for enhancing their readability with more clarity and sharability.

Regarding abbreviation resolution, we witnessed a number of the related works with many focuses on different tasks and purposes. As one of the first works considering medical abbreviations, [3] has provided a listing of medical abbreviations in 6 nonexclusive groups for English medical records. The result from [3] was well-known as Berman's list of abbreviations, helpful for abbreviation disambiguation in English clinical notes. Another effort in [23] has been given for normalizing abbreviations in clinical text and another one in [2] for enhancing the readability of discharge summaries. Furthermore, [21] has examined three natural language processing systems (MetaMap, MedLEE, cTAKES) to see how well these systems deal with abbreviations in English discharge summaries. The authors also suggested "accurate identification of clinical abbreviations is a challenging task". This suggestion is understandable because many abbreviations are dependent on context for their interpretation as mentioned in [12]. Indeed, [12] realized many abbreviations encountered have been commonly used but dependent on context. Thus, capturing the surrounding context of an abbreviation is important to distinguish itself from non-abbreviations in clinical text.

In this paper, we propose an effective solution to abbreviation identification in electronic medical records with level-wise feature engineering and supervised learning. Level-wise feature engineering is performed to characterize each token that is either an abbreviation or a non-abbreviation at the token, sentence, and note levels. Many aspects of each token (abbreviation or non-abbreviation) can be examined and captured to be able to discriminate between the tokens. Especially, their contexts are defined according to their surrounding neighbors in a continuous bag-of-words model introduced in [13]. As a result, a comprehensive vector representation for each token is achieved in a vector space. After that, many open options can be made to build an abbreviation identifier in a supervised learning mechanism. The resulting identifier can be used for identifying abbreviations automatically in clinical text. Experimental results on various real clinical note types have confirmed the effectiveness of our solution with high accuracy, precision, recall, and F-measure.

2 Abbreviation Identification in Electronic Medical Records with Level-wise Feature Engineering

In this section, we propose an abbreviation identification task on electronic medical records with level-wise feature engineering in the vector space. The task is defined in a broader view of the clinical note de-noising process with abbreviation resolution. It contributes to cleansing clinical texts of abbreviations, one kind of explicit noises.

2.1 De-noising Clinical Notes with Abbreviation Resolution

De-noising clinical notes with abbreviation resolution aims at replacing all abbreviations written in clinical notes with their correct long forms in order to improve the readability of these clinical notes and further enable their sharability for other medical and healthcare research activities. This abbreviation resolution consists of two phases: (1). Abbreviation Identification and (2). Abbreviation Disambiguation. The first phase needs to extract the parts from the free text in the clinical notes that are abbreviations and the second phase finds a long form corresponding to each abbreviation. As one long form might have many written abbreviations and one abbreviation might be used as a short form of many words or phrases, a correct single sense needs to be determined for an abbreviation and thus, abbreviation disambiguation is implied. These two phases are performed consecutively as shown in Fig. 1. The entire process will make clinical notes with noises (abbreviations) in their text cleansed and readable.

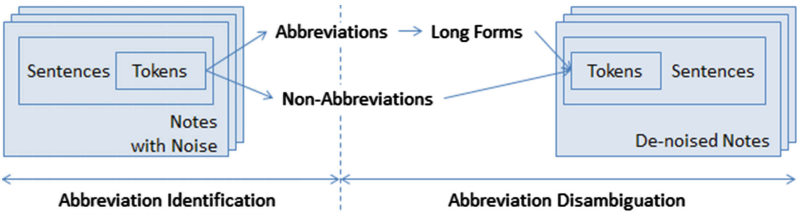

Fig. 1. De-noising clinical notes with abbreviation resolution

2.2 Abbreviation Identification Task Definition

In this subsection, we elaborate the abbreviation identification phase and define an abbreviation identification task.

As the input of this phase is a collection of clinical notes that contain free text and its output is a collection of abbreviations written in the clinical notes, we formulate an abbreviation identification task as a token-level binary classification task on the free text of the clinical notes. As well-known in data mining and machine learning, a classification task is performed in a supervised learning mechanism to classify given objects into several predefined classes. "Binary classification" means that there are two predefined classes (groups) corresponding to a group of abbreviations (class = 1) and another group of non-abbreviations (class = 0). "Token-level classification" means that each object in the classification task is defined at the token level of the clinical notes. Indeed, each object is a token obtained from the free text in the clinical notes. A token is either an abbreviation or a non-abbreviation. It is represented as a vector, called a token-level vector, so that a binary classification task can be performed in a vector space. A classification model of the binary classification task needs to be built for abbreviation identification and is called an abbreviation identifier used with the aforementioned input and producing the expected output. A token-level vector used in the phase of constructing the abbreviation identifier is called a training token-level vector. Each training vector is given a true class value which is either 1 for an

abbreviation or 0 for a non-abbreviation while a vector corresponding to a token that needs to be determined as an abbreviation or not will be assigned a class value after the model classifies its vector to an appropriate class. A vector in the vector space of the task that has p dimensions is characterized by p features corresponding to p dimensions of the vector space. A feature value is of any data type in implementation depending on feature engineering we perform to represent each token. In our work, each feature value is a real number and thus, a large number of supervised learning algorithms can be utilized for the task in a vector space.

2.3 Level-wise Feature Engineering

In order to represent each token in a vector space, we first design the structure of each token in the form of a vector and then process the clinical notes to generate a corresponding vector by extracting and calculating its feature values. In our work, we do feature engineering by capturing many different aspects of each token from the most detailed level to the coarsest one. In particular, we consider the features at the token, sentence, and note levels. That is why we call our feature engineering "level-wise feature engineering". It is delineated as follows.

At the ***token*** level, each token is characterized by its own aspects such as word form with orthographic properties, word length, and semantics. We use 3 orthographic features named *AnyDigit*, *AnySpecialChar*, and *AllConsonants*, using 1 word length feature named *Length*, and using 2 semantic features named *inDictionary* and *isAcronym*. These token-level features are described below:

- *AnyDigit*: indicating if the current token contains any digit such as "0", "1", "2", "3", "4", "5", "6", "7", "8", and "9". If yes, one (1) is the corresponding feature value. Otherwise, zero (0) is used. The use of digits in abbreviations is little; however, they might be used to shorten the long form of some number or combined with letters in abbreviations.
- *AnySpecialChar*: indicating if the current token contains any special character such as ".", ";", ",", "-", "_", "(", ")", "@", "%", "&", and so on. If yes, one (1) is the corresponding feature value. Otherwise, zero (0) is used. It is found that abbreviations don't often contain special characters except for "-", "_", and "." for connecting the components of their long forms.
- *AllConsonants*: indicating if the current token is composed of all consonants such as "b", "c", "d", …, "w", "x", "z". If yes, one (1) is the corresponding feature value. Otherwise, zero (0) is used. In our work, we consider acronyms to be special abbreviations. Thus, all abbreviations which are acronyms created by a sequence of the first letters of the components of their long forms tend to contain all consonants. Nonetheless, there exist some abbreviations including vowels.
- *Length*: the number of characters in the current token. It is found that most abbreviations are short due to time saving, the main purpose of abbreviation writing.
- *inDictionary*: indicating if the current token is included in a given medical dictionary. This dictionary is regarded as an external resource to provide us with the

semantics of the tokens in case they are medical terms in the bio-medical domain. If yes, one (1) is the corresponding feature value. Otherwise, zero (0) is used. This token-level feature helps realizing that the current token might be a non-abbreviation as medical terms in the dictionary are in their full forms.

- *isAcronym*: indicating if the current token matches any acronym of a medical term in the aforementioned dictionary. If yes, one (1) is the corresponding feature value. Otherwise, zero (0) is used. In contrast to *inDictionary*, this token-level feature helps us point out that the current token might be an abbreviation.

At the ***sentence*** level, many contextual features are defined from the surrounding words of each token in the sentence where it is contained. Different from the existing work [26] that used word forms of the surrounding words and [22] that used the characteristics of the previous/next word of each current word for capturing the context of the current word, we use a word embedding vector to encode the context of each token in a vector space. As introduced in [13], each word can be represented as a continuous vector in a vector space using either a continuous bag-of-words model or a continuous skip-gram model. The previous model predicts the current word based on the context words while the latter predicts the words in a certain range before and after the current word which is now an input. Due to the nature of our abbreviation identification task, which concentrates on deciding if a current word (token) is an abbreviation or not, we would like to capture the context of each token based on its surrounding words and thus in some certain contexts, long forms are not written, i.e. abbreviations are preferred and written. Therefore, a continuous bag-of-words model is an appropriate choice of focusing on the current token and generating the sentence-level contextual features. The number of the resulting sentence-level contextual features is the output layer size V of the continuous bag-of-words model.

At the ***note*** level, occurrence of each token in clinical notes is considered as a note-level feature. We use a term frequency to capture the number of occurrences of each token. Our feature engineering does not compute the percentage as we plan to perform our abbreviation resolution over time. As of that moment, updating term frequencies which are the number of occurrences is simpler than updating the percentage of each token because only the tokens in the incremental part need to be checked.

As a result, a token is represented in the following form of a vector:

$$\mathrm{X} = \left(\mathrm{x}^{\mathrm{t}}_1, \ldots, \mathrm{x}^{\mathrm{t}}_{tp}, \mathrm{x}^{\mathrm{s}}_1, \ldots, \mathrm{x}^{\mathrm{s}}_{sp}, \mathrm{x}^{\mathrm{n}}_1, \ldots, \mathrm{x}^{\mathrm{n}}_{np}\right)$$

in a vector space of $(tp + sp + np)$ dimensions where $\mathrm{X}^{\mathrm{t}}_i$ is a feature value of the i-th feature at the token level, $\mathrm{X}^{\mathrm{s}}_j$ a feature value of the j-th feature at the sentence level, and $\mathrm{X}^{\mathrm{n}}_k$ a feature value of the k-th feature at the note level; and tp is the number of token-level features, sp the number of sentence-level features, and np the number of note-level features. For our encodings in the abbreviation identification task, we design each token representation in a (7 + V)-dimension space with $tp = 6$, $sp = \mathrm{V}$, and $np = 1$ where V is sp, an output layer size in the continuous bag-of-words model.

2.4 Discussion

With the abbreviation identification task definition and level-wise feature engineering, the advantages of our level-wise feature engineering are highlighted:

Firstly, unlike the related works [14, 23, 24] where feature engineering is supervised with class information in terms of "target abbreviation", ours does level-wise feature engineering in an unsupervised manner with no class information. Thus, our approach is more practical and applicable for abbreviation identification in abbreviation resolution to the coming electronic medical records over time.

Secondly, our work has defined a comprehensive representation of each token in clinical notes that can capture many different aspects of each token from the most detailed level to the roughest one, suitable for the context of abbreviation usage where an abbreviation writing habit is formed, agreed, and maintained in a group of people. Thus, sentence- and note-level features get important for abbreviation determination while token-level features are remained to characterize each token.

Thirdly, there is no restriction on abbreviation writing styles as our feature engineering does not make use of abbreviation writing styles. Above all, our level-wise feature engineering has no restriction on note structures as only token occurrence is examined at the note level. Specific note structures were not encoded into the resulting features so that many various clinical types could be supported over time. Especially, we simply consider two groups of tokens: one for abbreviations and another one for non-abbreviations. This means that there is no distinguishing between abbreviations and acronyms, leading to the capability to identify a large set of so-called short forms, i.e. abbreviations, in clinical notes.

However, we would like to note that each group of the features obtained at each level in our level-wise feature engineering is not automatically and specifically selected for any given collection of clinical notes. We choose and design them level by level based on the nature of the abbreviation resolution task, our heuristic rules, and the existing features used in the related works [22–24, 26] mentioned above.

3 An Abbreviation Identification Process Using a Supervised Learning Mechanism on Electronic Medical Records

Based on the abbreviation identification task and level-wise feature engineering defined previously, an abbreviation identification process using a supervised learning mechanism on electronic medical records is proposed. Sketched in Fig. 2, our abbreviation identification process is conducted with three parts executed sequentially: A. Data Preparation; B. Identifier Construction; and C. Abbreviation Identification.

A. Data Preparation

As we transform an abbreviation identification task on free text in clinical notes into a classification task on token-level vectors in a vector space, data preparation plays an important role to generate appropriate token-level vectors from the given clinical notes. All clinical notes need to be gone through natural language processing as they contain free text. In order to gather a collection of tokens in these clinical notes, tokenization is

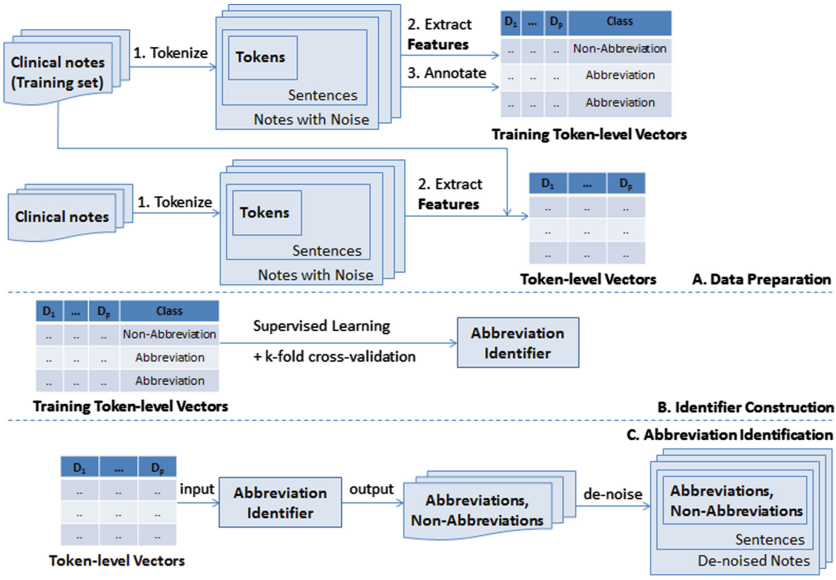

Fig. 2. The proposed abbreviation identification process using a supervised learning mechanism on electronic medical records

performed. For each resulting token, a vector is created as introduced earlier. In addition to tokenization, we use our proposed level-wise feature engineering for both training clinical notes and clinical notes that need abbreviation resolution. For both training token-level vectors and other token-level vectors, all token-level feature values are achieved after characterizing each token. All sentence- and note-level feature values are derived in an unsupervised approach after a continuous bag-of-words model is built and term frequencies are obtained using all clinical notes. An additional work is done for the tokens in the training set in order to get their class values by annotation. Annotating each token as either an abbreviation or a non-abbreviation is time-consuming but inevitable for a classification task. A true class value which is 1 or 0 is assigned to a true abbreviation or non-abbreviation in the training set as our classification task is a binary one.

An expected result of this part includes one collection of training token-level vectors and another collection of token-level vectors corresponding to the tokens in the clinical notes that are going to be de-noised with abbreviation resolution. Such an output will be fed into the next two parts for abbreviation identifier construction and abbreviation identification, respectively.

B. Identifier Construction

In our current work, the abbreviation identification task is carried out in a supervised learning mechanism as a classification task and thus, a collection of training token-level vectors needs to be ready for identifier construction in this part. After that, supervised learning is used to process these training vectors to return a classifier which is our

abbreviation identifier. Before this abbreviation identifier is shifted to the next part for use, an evaluation is made in a k-fold cross-validation where our level-wise feature engineering is guaranteed for the vectors in the training folds and the test fold in each loop. If its performance is not satisfied for the abbreviation identification task, supervised learning algorithms along with Data Preparation part and particularly our level-wise feature engineering need examining for more improvement.

As the training token-level vectors are formed in a conventional vector space, we can make the most of a large number of supervised learning algorithms which are available to support this part for identifier construction. This implies that our solution is practical and applicable to de-noising clinical notes in electronic medical records to enhance their readability.

C. Abbreviation Identification

This part comes after the previous two parts as soon as token-level vectors are transferred from Data Preparation part and an abbreviation identifier is sent by Identifier Construction part. It will make use of the abbreviation identifier to decide which token is an abbreviation and which is not by classifying a corresponding token-level vector. The resulting abbreviations and non-abbreviations are then marked for the corresponding tokens in the clinical notes so that long forms of the significant abbreviations can be resolved in the next phase. How correctly a token is marked with abbreviation or non-abbreviation relies on how effective the abbreviation identifier is. Thus, active learning with human interaction should be considered to further check the returned abbreviations and non-abbreviations in practice in the future.

D. Discussion

Regarding a theoretical evaluation of our solution, which is processed in a supervised learning mechanism with level-wise feature engineering, we would like to discuss its strengths and weakness as follows.

The first remarkable point of the proposed solution is that it spends more effort on data preparation and less effort on both identifier construction and abbreviation identification. This leads to more open options to identifier construction for efficiency and effectiveness in a supervised learning phase. Besides, our solution can make abbreviation identification convenient and applicable to clinical text in a classification phase. All abbreviations in clinical text can be identified automatically.

Another advantage of our solution is that it is general, i.e. not specific, for languages, note structures, and note types used for clinical text from feature engineering to process implementation. As a derived advantage, our solution is simple and practical to be deployed in practice and lay the basis for abbreviation resolution.

It is important for us to emphasize only one weakness of our solution regarding the synchronization between the feature extraction of training token-level vectors and other token-level vectors that need abbreviation identification. As our level-wise feature engineering is done not only at the token level but also at the sentence and note levels, all clinical notes have to be gathered together to build a continuous bag-of-words model and term frequencies in an unsupervised manner. Whenever there are new clinical notes for abbreviation identification, the feature extraction of all token-level vectors needs to be performed simultaneously with the same continuous bag-of-words

model and checking the same term frequencies so that sentence-level and note-level characteristics can be included in every token-level vector. As of that moment, a continuous bag-of-words model and term frequencies need to be re-generated. Nonetheless, such a weakness can be covered as we improve our solution soon in a semi-supervised learning mechanism over time.

4 Experimental Results

In order to evaluate the proposed solution, we present several experiments and provide discussions about their results. The experiments were conducted using the program written in Java for feature extraction, the word embedding implementation in Word2VecJava [20], and the supervised learning algorithm implementation in Weka 3 [18]. As an external resource, a hand-coded dictionary composed of 1995 medical terms in either Vietnamese or English is used in our experiments. Using this dictionary, we generated an acronym for each medical term regardless of its language. In the following, we elaborate our experiments for an evaluation using a triangle rule which is "at least three" for data, algorithms, and measures.

Clinical notes used in our experiments come from electronic medical records at Van Don Hospital in Vietnam [1], written in Vietnamese and English for medical terms. Details about clinical notes and abbreviations written in those notes are given in Table 1. Different from the related works, our work does not perform the identification task for each abbreviation as there are many distinct abbreviations in the current notes and the future ones. We also have no restriction on the minimum and maximum lengths of each abbreviation that needs to be identified. Therefore, for the identification task, we simply annotate each token as an abbreviation or a non-abbreviation. The percentage of abbreviations in each set of notes is calculated with respect to the total number of tokens in each set. Table 1 shows that there is an imbalance in our data set because the number of non-abbreviations is very high. This leads to our choice of evaluation measures which are Accuracy, Precision, Recall, and F-measure. Accuracy is used to check the overall performance of the resulting identifier whereas Precision, Recall, and F-measure are used for only the class of abbreviations to see how correctly the resulting identifier can recognize abbreviations. Furthermore, we prepare three different types of clinical notes including care notes, treatment order notes, and treatment progress notes. The note types are different from each other in the number of records, the number of sentences, the number of tokens, and the number of abbreviations.

Table 1. Details about clinical notes and abbreviations

Clinical note types	Care	Treatment order	Treatment progress
Patient#	2,000	2,000	2,000
Record#	12,100	4,175	4,175
Sentence#	8,978	39,206	13,852
Token#	52,109	325,496	138,602
Abbreviation#	3,031	24,693	7,641
Abbreviation%	5.82	7.59	5.51

Regarding supervised learning algorithms, we apply Random Forest with 100 trees, C4.5, and k-NN with k = 1 and Euclidean metric. The selected algorithms belong to a various number of the supervised learning categories because Random Forest is an ensemble method widely-used in the related works [22], C4.5 is a baseline decision tree algorithm also used in the related works [22, 25], and k-NN follows a lazy learning approach. Moreover, these three different algorithms have been studied very well in the machine learning research area and thus we do not focus on the fact that the best results are from which algorithm. Instead, we concentrate on a common response from these algorithms for the results of the task in our solution when several various feature sets were extracted at three different levels of detail.

In the following tables, experimental results are displayed. The experimental results on care notes are given in Table 2, the one on treatment order notes in Table 3, and the one on treatment progress notes in Table 4. In addition to these tables, we show the experimental results with different layer sizes for the continuous bag-of-word models using Random Forest in Table 5.

Table 2. Experimental results on care notes

Algorithm	Measure	Token-level	Token-level + Note-level	Sentence-level - 5	Combination-5	Combination with no external resource - 5
Random Forest	Accuracy	98.87	99.948	99.95	99.987	**99.989**
	Precision	96.1	99.8	**100**	**100**	**100**
	Recall	84	99.3	99.2	**99.8**	**99.8**
	F-measure	89.6	99.5	99.6	**99.9**	**99.9**
C4.5	Accuracy	98.879	99.931	99.86	99.941	**99.946**
	Precision	96.3	99.7	99.2	**99.8**	**99.8**
	Recall	83.9	99.1	98.4	**99.2**	**99.2**
	F-measure	89.7	99.4	98.8	**99.5**	**99.5**
1-NN	Accuracy	98.877	99.948	99.916	99.96	**99.964**
	Precision	96.3	**99.8**	99.3	99.7	99.7
	Recall	83.9	99.3	99.2	99.6	**99.7**
	F-measure	89.7	99.5	99.2	99.6	**99.7**

From the results in Tables 2, 3 and 4, we realize that sentence-level features with layer size = 5 can help achieving higher precision while token-level and note-level features achieving higher recall. In general, a combination of the features at all levels with layer size = 5 gets the highest accuracy in most cases, regardless of the selected supervised learning algorithm. Besides, we found little difference between using and not using any external resource. This might be because as of this moment, our hand-coded dictionary has a limited number of terms. In our opinion, semantic features are important to reflect each token and especially potential for the next phase to disambiguate the senses of each abbreviation.

Table 3. Experimental results on treatment order notes

Algorithm	Measure	Token-level	Token-level + Note-level	Sentence-level - 5	Combination-5	Combination with no external resource - 5
Random Forest	Accuracy	96.746	99.603	99.664	**99.672**	99.671
	Precision	81.1	97.8	**98.2**	**98.2**	**98.2**
	Recall	74.4	96.9	97.4	**97.5**	**97.5**
	F-measure	77.6	97.3	**97.8**	**97.8**	**97.8**
C4.5	Accuracy	96.747	99.601	99.624	**99.66**	99.659
	Precision	81.1	97.8	97.9	**98.1**	**98.1**
	Recall	74.4	96.9	97.1	**97.4**	**97.4**
	F-measure	77.6	97.3	97.5	**97.7**	**97.7**
1-NN	Accuracy	96.746	99.603	99.657	99.668	**99.669**
	Precision	81.1	97.8	98.1	**98.2**	**98.2**
	Recall	74.4	96.9	97.4	**97.5**	**97.5**
	F-measure	77.6	97.3	97.7	**97.8**	**97.8**

Table 4. Experimental results on treatment progress notes

Algorithm	Measure	Token-level	Token-level + Note-level	Sentence-level - 5	Combination-5	Combination with no external resource - 5
Random Forest	Accuracy	97.509	99.789	99.895	**99.907**	99.903
	Precision	94	98.5	**99.8**	99.5	99.4
	Recall	58.6	97.7	98.3	**98.8**	**98.8**
	F-measure	72.2	98.1	99	**99.1**	**99.1**
C4.5	Accuracy	97.51	99.789	99.782	99.864	**99.867**
	Precision	94	98.3	98.8	99	**99.1**
	Recall	58.6	97.9	97.3	**98.5**	**98.5**
	F-measure	72.2	98.1	98	98.7	**98.8**
1-NN	Accuracy	97.509	99.789	99.848	**99.888**	**99.888**
	Precision	94	98.5	98.8	**99**	**99**
	Recall	58.6	97.6	98.5	**99**	98.9
	F-measure	89.7	99.5	99.2	99.6	**99.7**

Table 5. Experimental results with different layer sizes using random forests

Note type	Layer size	Sentence-level				Combination with an external resource			
		Accuracy	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure
Care Notes	5	99.950	**100.0**	99.2	99.6	**99.987**	**100.0**	**99.8**	**99.9**
	100	99.950	**100.0**	99.1	99.5	99.952	**100.0**	99.2	99.6
Treatment Order Notes	5	99.664	**98.2**	97.4	**97.8**	**99.672**	**98.2**	**97.5**	**97.8**
	100	99.665	**98.2**	97.4	**97.8**	99.665	**98.2**	97.4	**97.8**
Treatment Progress Notes	5	99.895	**99.8**	98.3	99.0	**99.907**	99.5	**98.8**	**99.1**
	100	99.897	**99.8**	98.3	99.0	99.899	**99.8**	**98.4**	**99.1**

From the results in Table 5, contextual features extracted by means of continuous bag-of-word models play an important role in correctly recognizing an abbreviation with very high precision but with acceptable recall. As we extend the vector space with other features at token and note levels, the abbreviation identifier can be enhanced with higher recall while precision appears to be remained. Hence, its final accuracy can be improved. This point is highlighted for both layer sizes which are 5 and 100 and for three different note types. Furthermore, it is worth noting that for these clinical notes, five contextual features extracted for each token are good enough for abbreviation identification as the experimental results with layer size = 5 are often better than those with layer size = 100. Nonetheless, "which layer size is suitable for abbreviation identification on which notes" is open in the future for the task as our work has not yet determined an appropriate value for this parameter automatically. With 5 contextual features at the sentence level, the cost for our constructing the abbreviation identifier gets increased a little as compared to that cost with 50 or 100 word embedding features in the existing works such as [24] and [11], respectively. Therefore, the total number of the resulting features at all levels in our Combination-5 experiments is 12, small but effective for abbreviation identification in those notes.

In short, our work has provided an effective solution to abbreviation identification that can be used with many various existing supervised learning algorithms and a comprehensive representation of each token in clinical notes. This solution has been examined on the real clinical notes and produced promising results on a consistent basis. It is then able to lay the foundations for abbreviation resolution in the next step where long forms need to be decided for each correctly identified abbreviation.

5 Related Works

For a comparison between the existing works and ours, an overall review on the related works is given in this section. We also present the rationales behind our experiments, which did not include any empirical comparison with these works.

First of all, we give a general review on the works in [2, 5–7, 9–11, 14–17, 19, 22–26] related to abbreviation resolution. Among these works, many related tasks have been considered to make some certain contribution to abbreviation resolution in clinical text and then further to noise cleaning and readability improvement on the clinical text in electronic medical records. For example, [10, 22, 25] focused on abbreviation detection, [5–7, 9, 11, 14, 15, 17, 19, 23, 24] on abbreviation disambiguation and expansion, and [11, 16, 24–26] on sense inventory construction for abbreviations. At this moment, our work concentrates on the first important phase of abbreviation resolution that is abbreviation identification. Using the abbreviations correctly identified, long forms can be determined and assigned to each abbreviation.

Secondly, we would like to discuss the reasons for not comparing the related works with ours in the experiments presented previously. As discussed earlier, our work aims to a more general solution to abbreviation identification. In contrast, a few related works were specific for dealing with some kinds of abbreviations in clinical text. For example, [10] has connected their solution to German abbreviation writing styles and [9] has paid attention to only the abbreviations that are 3 letters long. Another

important point is that our work follows an unsupervised approach to level-wise feature engineering to be able to handle other unknown abbreviations in the future. Different from ours, the related works in [14, 23, 24] used a supervised approach in their feature engineering with respect to "target abbreviations". Our work also captures the context of each token at the sentence level using a continuous bag-of-words model in a vector space while [22] only uses local context based on the characteristics of the previous/next word of each current word and [26] uses word forms of the surrounding words. Besides, our work defines a comprehensive token-level vector representation with level-wise features in a vector space while many machine learning-based related works such as [14, 15, 25] are not based on a vector space model, leading to the different representations for clinical notes. Moreover, our work is based on a supervised learning mechanism for abbreviation identifier construction while several related works have made use of regular expressions [11], word lists and heuristic rules [25] for abbreviation identification. Last but not least, there is no available benchmark clinical data set for abbreviation identification in the present for empirical comparisons to be made with no bias. Each work has resolved this task using its own data set perhaps because of a high cost for data preparation, especially in machine learning-based works requiring large annotated data sets.

Based on our differences from the related works, it can be seen that our work has the merits of abbreviation identification so that a correct set of abbreviations can be further taken into consideration for finding their true long forms and de-noising clinical notes for readability and sharability enhancement. It has been proved with the high effectiveness of the proposed solution on various real clinical note types.

6 Conclusion

In this paper, we have taken into account an abbreviation identification task on electronic medical records. This task is formulated as a binary classification task to handle the first phase of abbreviation resolution to make electronic medical records more readable and sharable with long forms of their abbreviations. In our solution, we do level-wise feature engineering to represent each token in clinical notes in a vector space using several different aspects at token, sentence, and note levels corresponding to orthographic, word length, and semantic features at the token level; contextual features at the sentence level using the continuous bag-of-word model; and occurrence feature of each token in a given note set at the note level using term frequency, respectively. Many various existing supervised learning algorithms are then able to be utilized with the resulting token-level vectors to build an abbreviation identifier. We believe that a comprehensive set of level-wise features can help us distinguish instances of abbreviations from the others of non-abbreviations. Furthermore, our feature set does not rely much on external resources for semantics and the structure of each note type. In addition, it is proved with experimental results on real Vietnamese clinical data sets of three note types that our solution is really effective with very high accuracy, precision, recall, and F-measure values. This implies that abbreviation identification is tackled well for de-noising clinical notes with abbreviation resolution.

In the future, we plan to make our current solution more practical over time by using a semi-supervised learning mechanism instead of a supervised learning one. Besides, cleaning abbreviations from clinical notes by determining their correct long forms is one of our next steps to prepare electronic medical records for readability and sharability in further data analysis and knowledge discovery. Parallel processing for our solution to abbreviation resolution is also regarded to speed up the task. Finally, we will pay more attention to automatically determining parameter values in the task.

Acknowledgments. This work is funded by Vietnam National University at Ho Chi Minh City under the grant number B2016-42-01. In addition, we would like to thank John von Neumann Institute, Vietnam National University at Ho Chi Minh City, very much to provide us with a very powerful server machine to carry out the experiments. Moreover, this work was completed when the authors were working at Vietnam Institute for Advanced Study in Mathematics, Vietnam. Besides, our thanks go to Dr. Nguyen Thi Minh Huyen and her team at University of Science, Vietnam National University, Hanoi, Vietnam, for external resources used in the experiments and also to the administrative board at Van Don Hospital for their real clinical data and support.

References

1. A Set of Electronic Medical Records, Van Don Hospital, Vietnam, 24 February 2016
2. Adnan, M., Warren, J., Orr, M.: Iterative refinement of SemLink to enhance patient readability of discharge summaries. In: Grain, H., Schaper, L.K. (eds.) Health Informatics: Digital Health Service Delivery - The Future is Now!, pp. 128–134 (2013)
3. Berman, J.J.: Pathology abbreviated: a long review of short terms. Arch. Pathol. Lab. Med. **128**, 347–352 (2004)
4. Collard, B., Royal, A.: The use of abbreviations in surgical note keeping. Ann. Med. Surg. **4**, 100–102 (2015)
5. Henriksson, A., Moen, H., Skeppstedt, M., Daudaravičius, V., Duneld, M.: Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J. Biomed. Semant. **5**(6), 1–25 (2014)
6. Kim, Y., Hurdle, J., Meystre, S.M.: Using UMLS lexical resources to disambiguate abbreviations in clinical text. In: AMIA Annual Symposium Proceedings, pp. 715–722 (2011)
7. Kim, J.-B., Oh, H.-S., Nam, S.-S., Myaeng, S.-H.: Using candidate exploration and ranking for abbreviation resolution in clinical documents. In: Proceedings of the 2013 International Conference on Healthcare Informatics, pp. 317–326 (2013)
8. Kim, M.-Y., Xu, Y., Zaiane, O.R., Goebel, R.: Recognition of patient-related named entities in noisy tele-health texts. ACM Trans. Intell. Syst. Technol. **6**(4), 59:1–59:23 (2015)
9. Kim, S., Yoon, J.: Link-topic model for biomedical abbreviation disambiguation. J. Biomed. Inform. **53**, 367–380 (2015)
10. Kreuzthaler, M., Schulz, S.: Detection of sentence boundaries and abbreviations in clinical narratives. BMC Med. Inform. Decis. Making **15**, 1–13 (2015)
11. Liu, Y., Ge, T., Mathews, K.S., Ji, H., McGuinness, D.L.: Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing, pp. 92–97 (2015)
12. Long, W.J.: Parsing free text nursing notes. In: AMIA Annual Symposium Proceedings, p. 917 (2003)

13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Workshop Proceedings of the International Conference on Learning Representations (2013)
14. Moon, S., Berster, B.T., Xu, H., Cohen, T.: Word sense disambiguation of clinical abbreviations with hyperdimensional computing. In: AMIA Annual Symposium Proceedings, pp. 1007–1016 (2013)
15. Moon, S., McInnes, B., Melton, G.B.: Challenges and practical approaches with word sense disambiguation of acronyms and abbreviations in the clinical domain. Healthc. Inform. Res. **21**(1), 35–42 (2015)
16. Moon, S., Pakhomov, S., Liu, N., Ryan, J.O., Melton, G.M.: A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. J. Am. Med. Inform. Assoc. **21**, 299–307 (2014)
17. Pakhomov, S., Pedersen, T., Chute, C.G.: Abbreviation and acronym disambiguation in clinical discourse. In: AMIA Annual Symposium Proceedings, pp. 589–593 (2005)
18. Weka 3, Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka. Accessed on 22 February 2016
19. Wong, W., Glance, D.: Statistical semantic and clinician confidence analysis for correcting abbreviations and spelling errors in clinical progress notes. Artif. Intell. Med. **53**, 171–180 (2011)
20. Word2VecJava. https://github.com/medallia/Word2VecJava. Accessed on 22 February 2016
21. Wu, Y., Denny, J.C., Rosenbloom, S.T., Miller, R.A., Giuse, D.A., Xu, H.: A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In: AMIA Annual Symposium Proceedings, pp. 997–1003 (2012)
22. Wu, Y., Rosenbloom, S.T., Denny, J.C., Miller, R.A., Mani, S., Giuse, D.A., Xu, H.: Detecting abbreviations in discharge summaries using machine learning methods. In: AMIA Annual Symposium Proceedings, pp. 1541–1549 (2011)
23. Wu, Y., Tang, B., Jiang, M., Moon, S., Denny, J.C., Xu, H.: Clinical acronym/abbreviation normalization using a hybrid approach. In: CLEF (2013)
24. Wu, Y., Xu, J., Zhang, Y., Xu, H.: Clinical abbreviation disambiguation using neural word embeddings. In: Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pp. 171–176 (2015)
25. Xu, H., Stetson, P.D., Friedman, C.: A study of abbreviations in clinical notes. In: AMIA Annual Symposium Proceedings, pp. 822–825 (2007)
26. Xu, H., Stetson, P.D., Friedman, C.: Methods for building sense inventories of abbreviations in clinical notes. J. Am. Med. Inform. Assoc. **16**(1), 103–108 (2009)