EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 150 No. 10
October 2021

Research in Computing Science

RESEARCH IN COMPUTING SCIENCE, *Año 21, Volumen 150, No. 10, 08 de Octubre de 2021, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203 ISSN: en trámite, ambos otorgados por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr.Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 08 de Octubre de 2021.*

RESEARCH IN COMPUTING SCIENCE, *Year 21, Volume 150, No. 10, October 08, 2021, is a monthly publication edited by the National Polytechnic Institute through the Center for Computing Research. Av. Juan de Dios Bátiz S/N, Esq. Miguel Othón de Mendizábal, Nueva Industrial Vallejo, C.P. 07738, Mexico City, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor in charge: Dr. Grigori Sidorov. Reservation of Exclusive Use Rights of Title No. 04-2019-082310242100-203. ISSN: pending, both granted by the National Copyright Institute. Responsible for the latest update of this issue: the Computer Research Center, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Last modified on October 08, 2021.*

Volume 150(10)

Advances in Artificial Intelligence

Miguel González Mendoza
Juan Martínez Miranda (eds.)

Instituto Politécnico Nacional,
Centro de Investigación en Computación, México 2021

ISSN: en trámite

Formerly ISSNs: 1870-4069, 1665-9899

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Towards an English Reading Intelligent Tutor System for Native Spanish-Speaking Undergraduate Students

Adelina Escobar Acevedo, Josefina Guerrero García

Meritorious Autonomous University of Puebla,
Mexico

adeesa32@gmail.com

Abstract. Intelligent Tutor Systems (ITS) are challenging in domains as reading and writing, and target users require specific material to mature reading abilities. We are developing a methodology to produce reading ITS for undergraduate students, facing decisions in both fields, educational and computational; the contribution is to provide suitable material of interest to each user among available material.

Keywords: Intelligent tutoring systems, reading comprehensions, readability metrics.

1 Motivation

Reading is essential for school and life. In the international standard test PISA, Mexico is statistically significantly below average scores [1], undergraduate students do not perform better. Reading in a foreign language adds challenges to the task. Educational technologies may be helpful out of the classroom to improve students' skills. Advantages include following a personalized path, working without teachers' time constraints, spending more time on topics of interest or need reinforcement. Intelligent Tutoring Systems (ITS) have proved helpful for many disciplines, mainly in well-defined domains such as mathematics where there is only one possible answer, ill-defined domains like reading and writing are still challenging [2].

2 Previous Work in the Area

Reading ITS are scarce. They are based on their own corpus and targeted for specific users as the material is different for each. iSTART [3] is the closest to our project as it was developed for high school and college students. CAERS [4], ITSS [5], and EMBRACE [6] were developed for children. AutoTutor [7] and iSTART-All [8] for low literacy adults. Further information on art state refers to [9]. Only AutoTutor allows the user some freedom on the material to use.

Related advances that could be included in a reading ITS are text simplification [10], automatic readability metrics, overlapped clustering [11], question generation [12],

question classification [13], automatic essay evaluation [14], interpretable semantic similarities [15], among others.

3 Research Objectives

If the actual development of Natural Language Processing is enough, it is possible to follow a methodology to create reading ITS for English as Second Language (ESL) undergraduate students. The objective of the work is "To design and evaluate a methodology to develop an ITS for reading in English as Second Language for Spanish-Speaking undergraduate students". Specific objectives are:

- Gather reading material in English and classify them to conform domain module.
- Integrate activities pre and post-reading to develop reading comprehension abilities.
- Create a model and build an ITS prototype.
- Validate the ITS among ESL expert teachers.

4 Methodology

Every ITS requires enough material to provide to the student. Natural language processing requires corpus, a group of texts. Corpora are expensive to create in time and effort as it is manually reviewed or tagged.

The first step requires identifying text complexity, by language level or by the obtention of Flesh-Kinkaid and RDL2 readability metrics on every text in the corpus. Previous works have empirically proven that RDL2 is useful in second language learners' reading tasks.

The second phase, to recommend texts to readers, requires obtaining documents similarity. The algorithm should provide at least one similar document to the last read by the user and some others very different from it. As there are no right or wrong suggestions, the idea is to form clusters to represent groups of topics of interest. From those, the algorithm will suggest texts for the user's reading level.

The third phase needs performance evaluation, which is a vast area in reading comprehension. Three main methods consist of questions, writing, and diagram development. The first one is popular among second language learners; the second option is more common in monolingual students as it requires additional mind processes to produce text. The third option is used to facilitate information visualization mainly for young readers. All types of evaluation must be automatic in the ITS. First experiments are intended to include human elaborated questions attached to texts.

User experiments will include students with access to the system and a control group without the system. From the experiment, we will observe the performance of both groups in the reading comprehension assessments and apply a survey on interest. The idea is to provide at least five sessions to both groups and gather information related to reading performance in Spanish and English.

Among tools considered, Coh-Metrix for metrics and other text scores. Natural Language Toolkit NLTK [16] and Stanza [17]. Python as the programming language.

5 State of the Research

First, a small expert group was created, and the four members are teachers experienced in English reading comprehension from different Mexican universities. Second, a survey was developed to register English teachers' experience in classrooms about reading. Their observations allowed us to set goals on comprehension, user freedom, and strategies.

Tutor architecture was adopted from [18], defining domain, tutor, student, and interface modules. Also, some ideas from English websites were collected to design a scalable navigation map.

The first module to create is the domain, as it contains the materials. We started creating a corpus from pages for English Second Language readers, but we have not collected enough material targeted for adults with creative commons license and reading comprehension questions. While that corpus grows, we are working with the OneStopEnglish corpus [19]. The corpus allows three manually simplified versions of each text and provides POS tags and common readability metrics, but it has no questions.

Some fast tests indicated that reading levels are not aligned with CEFR. Additional text analysis was performed with Coh-Metrix to obtain readability metrics for each text. Results show a brief comparison among them, clarifying the deepness of information used to obtain them. This analysis included the tool used to rank text in CEFR levels.

We have already started the first experiments to create topic groups, and we are adjusting parameters. In parallel, a basic recommendation system was constructed using cosine similarity. A pilot exercise is desired soon to obtain users' opinions and suggestions.

References

1. OCDE: Programa para la Evaluación Internacional de Alumnos (PISA) 2018 - Resultados México (2019)
2. Jacovina, M.E., McNamara, D.S.: Intelligent tutoring systems for literacy: Existing technologies and continuing challenges. In: Intelligent Tutoring Systems: Structure, Applications and Challenges. Nova Science Publishers Inc., pp. 153–74 (2016)
3. McCarthy, K.S., Watanabe, M., Dai, J., McNamara, D.S.: Personalized learning in iSTART: Past modifications and future design. J Res Technol Educ., 2, 52(3), pp. 301–321 (2020) doi: 10.1080/15391523.2020.1716201.
4. Chiang, K., Fan, C., Liu, H., Chen, G.: Effects of a computer-assisted argument map learning strategy on sixth-grade students' argumentative essay reading comprehension. Multimed Tools Appl., 4, 75(16), pp. 9973–9990 (2016) doi: 10.1007/s11042-015-2904-y.
5. Wijekumar, K.K., Meyer, B.J.F., Lei, P., Cheng, W., Ji, X., Joshi, R.M.: Evidence of an Intelligent Tutoring System as a Mindtool to Promote Strategic Memory of Expository Texts and Comprehension With Children in Grades 4 and 5. J Educ Comput Res., 5, 55(7), pp. 1022–1048 (2017) doi: 10.1177/0735633117696909.
6. Walker, E., Wong, A., Fialko, S., Restrepo, M.A., Glenberg, A.M.: EMBRACE:

Applying Cognitive Tutor Principles to Reading Comprehension. In: International Conference on Artificial Intelligence in Education 2017. pp. 578–81 (2017). doi: 10.1007/978-3-319-61425-0_68.
7. Fang, Y., Lippert, A., Cai, Z., Hu, X., Graesser, A.C.: A Conversation-Based Intelligent Tutoring System Benefits Adult Readers with Low Literacy Skills. In: Lecture Notes in Computer Science, Springer International Publishing, pp. 604–14 (2019) doi: 10.1007/978-3-030-22341-0_47.
8. Johnson, A.M., Guerrero, T.A., Tighe, E.L., McNamara, D.S.: iSTART-ALL: Confronting Adult Low Literacy with Intelligent Tutoring for Reading Comprehension. In: Aied 2017, pp. 125–36 (2017) doi: 10.1007/978-3-319-61425-0-11.
9. Escobar-Acevedo, A., Guerrero-García, J.: Revisión de Tecnologías Educativas que Fomentan la Lectura de Comprensión Autónoma, Brazilian J Comput Educ., 29, pp. 980–992 (2021)
10. Al-Thanyyan, S.S., Azmi, A.M.: Automated Text Simplification: A survey. ACM Comput Surv., 54(2), pp. 1–36 (2021) doi:10.1145/3442695.
11. Beltrán, B., Vilariño, D., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Pinto D.: K-means based method for overlapping document clustering. J Intell Fuzzy Syst., 39(2), pp. 2127–2135 (2020)
12. Kurdi, G., Leo, J., Parsia, B., Sattler, U., Al-Emari, S.: A Systematic Review of Automatic Question Generation for Educational Purposes. Int J Artif Intell Educ. 21, 30(1), pp. 121–204 (2020) doi: 10.1007/s40593-019-00186-y.
13. Pad, U.: Question Difficulty – How to Estimate Without Norming, How to Use for Automated Grading, pp. 1–10 (2016)
14. Crossley, S., Allen, L.K.: Snow EL, McNamara DS. Pssst... textual features... there is more to automatic essay scoring than just you!. In: Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15. ACM Press, pp. 203–207 (2015) doi: 10.1145/2723576.2723595.
15. Majumder, G., Pakray, P., Das, R., Pinto, D.: Interpretable semantic textual similarity of sentences using alignment of chunks with classification and regression. Appl Intell., pp. 1–18 (2021)
16. Bird, S., Loper, E., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009) http://www.nltk.org/book_1ed/.
17. Qi, P., Zhang, Y., Zhang, Y., Bolton, J.: Manning CD. Stanza: A Python natural language processing toolkit for many human languages (2020) arXiv:2003.07082.
18. Cataldi, Z., Lage, F.J.: Sistemas tutores inteligentes orientados a la enseñanza para la comprensión. Revista Electrónica de Tecnología Educativa, pp. 1–19 (2009)
19. Vajjala, S., Lucic, I.: OneStopEnglish corpus : A new corpus for automatic readability assessment and text simplification. In: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, pp. 297–304 (2018)

ISSN 1870-4069

Enhancing Metacognitive Skills in Language Learning through a Conversational Pedagogical Agent

Joaquin Navarro Perales

Universidad Nacional Autónoma de México,
Mexico

Universidad Internacional de la Rioja,
Spain

joaquin_navarro@cuaieed.unam.mx

Abstract. Intelligent tutoring systems can adapt to learners to the extent of preventing them from making poor decisions related to underestimating or overestimating their abilities, however, they must be exposed to making poor decisions to develop their autonomy. This paper proposes the use of a conversational pedagogical agent to support language learners while solving grammar exercises, so that they can be aware of their performance level and strengthen their metacognitive skills. Students will interact with the agent through an instant messaging application and answer the Metacognitive Awareness Inventory (MAI) to identify whether there are significant differences in using or not using the agent. A qualitative evaluation of the learners' experience using the agent will also be conducted.

Keywords: Intelligent tutoring systems, conversational pedagogical agents, autonomous learning, adaptive learning, metacognition.

1 Motivation

Intelligent tutoring systems (ITSs) are learning environments that rely on the ability of intelligent algorithms to adapt to learners at a fine-grade level [1]. This can be done according to two adaptation techniques: adaptability and adaptivity. The first one allows the learners to take control, while the second empowers the system [2].

Some ITSs have a shared student/system control to prevent students from making suboptimal decisions and elicit high motivation [3]. However, when one of the goals is to promote learner autonomy, it is necessary to expose them to make suboptimal decisions so that they can correct them with the help of the system.

Another important fact is that previous research reports a lack of mobile-based ITSs. This is observed in two systematic reviews, one of which reported 54.71% of web-based systems and 15.09% mobile-based systems [4], while the other study found no systems capable of providing mobile-based content, even though mobile devices are considered emerging technologies and their use is part of daily routine [5].

According to the described above, this paper proposes an ITS based on a conversational agent that supports language learners to enhance their metacognitive

skills, i.e., skills related to planning, monitoring and evaluation, while solving grammar exercises. Different paths will be proposed including exercises corresponding to different mastery levels, and the learners will be able to decide which one to take, if they consider it necessary, they will be able to change levels. The system will be linked to the instant messaging app Telegram.

2 Previous Works in the Area

In [6], the teaching of Spanish punctuation through a chatbot was compared to the traditional method through written exercises on paper. A quasi-experimental design and a quantitative methodology were used during pre-test and post-test in a control and experimental group. Subsequently, the perception of the experimental group was analyzed through text mining in a forum. Their results showed that the students who used the conversational agent substantially improved their performance compared to the students who used the traditional method. In addition, the perception of the students who used the agent showed that they positively value it in three aspects: support and companionship in their learning process, greater feedback, and the possibility of interacting and learning anywhere and anytime.

A study comparing the use of two self-regulation strategies through an ITS is described in [7]. Participants speak English as their second language and performed exercises to improve their speaking ability according to three groups: one that used a help-seeking strategy, a second one using a self-monitoring strategy, and a control group that used the traditional teaching method that does not focus on a self-regulation strategy. A mixed method, quasi-experimental, pre-test, post-test design was used. The pre-test, and post-test are aligned with the International English Language Testing System (IELTS) speaking test. They comprise the quantitative section of the study. The qualitative section is constituted by a questionnaire about learning through an ITS. Their results revealed that self-monitoring strategy had a more significant effect on the students' performance than help-seeking. Students who used the ITS were satisfied with its capability to support them.

The systems described above are web-based, so it is important to mention that [8] used Rain Classroom, a mobile-supported tool based on the instant messaging app WeChat. In that study, the Critical Thinking Skills Survey (CTSS) was adopted to measure the critical thinking skills (CTS) of learners of English as a second language. A pre-test-post-test non-equivalent quasi-experimental design was applied to compare the CTS of learners instructed under a Rain-Classroom-based ITS with those taught by the traditional lecture approach. Their results indicated that the ITS had a positive effect on students' overall CTS, with significant improvement in the interpretation subscale, but had almost no influence on the evaluation and self-regulation subscales.

3 Research Objectives

The objective of this study is to verify if a conversational pedagogical agent helps English language learners enhance their metacognitive skills while practicing grammar exercises, compared to solving exercises without additional guidance.

4 Methodology

A mixed method, quasi-experimental, pre-test, post-test design will be used. Learners will take the Spanish version of the Metacognitive Awareness Inventory (MAI) [9] as the pre-test, and they will be divided in the experimental and control group. The experimental group will use a conversational agent as a support tool for solving English grammar exercises, while the control group will solve them using a Google form. Participants in both groups will have to select their level between basic, intermediate, and advanced, while the experimental group will have access to feedback and hints from the agent. After answering the exercises, the two groups will answer the MAI again. Finally, students in the experimental group will write a short text about their experience with the chatbot.

The conversational agent will be implemented on the Dialogflow platform and will be linked to the Telegram app through the BotFather framework. In this way, each participant in the experimental group will access a Telegram group in which the agent will appear as a member.

The quantitative phase covers independent sample t-tests to compare the mean scores and change scores of MAI results in both groups, and for the measurement of the effect size, Cohen's *d* will be used [8]. In the qualitative phase, a discourse analysis will be conducted with the texts on the students' experience.

5 State of the Research

The stage of defining the state-of-the-art on intelligent tutoring systems to improve metacognition has been completed. The selection of grammar exercises for the conversational agent is currently underway.

6 Contribution

This work contributes to the field of artificial intelligence in education by going against the mainstream trend of empowering algorithms to provide fully tailored solutions to learners, instead returning autonomy to learners and encouraging shared adaptation between learners and intelligent tutoring systems. This could lessen the gap between artificial intelligence and the humanities.

References

1. Graesser, A.C., Hu, X., Sottilare, R.: Intelligent Tutoring Systems. In: Fischer, F., Hmelo-Silver, C.E., Goldman, S.R., Reimann, P. (eds.) International Handbook of the Learning Sciences. Routledge (2018)
2. Dascalu, M.-I., Nitu, M., Alecu, G., Bodea, C.-N., Moldoveanu, A.D.B.: Formative Assessment Application with Social Media Integration Using Computer Adaptive Testing Techniques. In: Campbel, L.O. and Hartshorne, R. (eds.) Proceedings of the 12th International Conference on E-Learning, pp. 56–65 (2017)

3. Long, Y., Aleven, V.: Mastery-Oriented Shared Student/System Control Over Problem Selection in a Linear Equation Tutor. In: Micarelli, A., Stamper, J., and Panourgia, K. (eds.) Intelligent Tutoring Systems, ITS 2016, Lecture Notes in Computer Science, pp. 90–100 (2016)
4. Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, R., Rakhshan, S., Keikha, M., Ghazi Saeedi, L.: Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. Interactive Learning Environments, pp. 11–22 (2018)
5. Soofi, A.A., Uddin, M.: A Systematic Review of Domains, Techniques, Delivery Modes and Validation Methods for Intelligent Tutoring Systems. IJACSA, 10 (2019)
6. Vázquez-Cano, E., Mengual-Andrés, S., López-Meneses, E.: Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. Int J Educ Technol High Educ., 18, 33 (2021)
7. Mohammadzadeh, A., Sarkhosh, M.: The Effects of Self-Regulatory Learning through Computer-Assisted Intelligent Tutoring System on the Improvement of EFL Learners' Speaking Ability. International Journal of Instruction, 11, pp. 167–184 (2018)
8. Chen, J., Hu, J.: Enhancing L2 Learners' Critical Thinking Skills Through a Connectivism-Based Intelligent Learning System. International Journal of English Linguistics, 8, pp. 12–21 (2018)
9. Huertas Bustos, A.P., Vesga Bravo, G.J., Galindo León, M.: Validación del instrumento 'inventario de habilidades metacognitivas (Mai)' con estudiantes colombianos. Prax. Saber, 5, pp. 55 (2014)

ISSN 1870-4069

Environmental Monitoring for Estimation of Indoor Occupancy Levels based on Unsupervised Machine Learning for COVID-19 Restrictions

Alma Rosa Mena Martínez, Héctor Ceballos, Joanna Alvarado-Uribe

Instituto Tecnológico y de Estudios Superiores de Monterrey, Mexico

a00834070@tec.mx,ceballos@tec.mx,joanna.alvarado@tec.mx

Abstract. Today, the COVID-19 pandemic has imposed stricter regulations on social distancing and crowd control, these policies can be managed effectively through the occupancy information of the place. Nevertheless, to predict the number of occupants accurately, there needs to be sufficient labeled data for training the predictors and validating the model. Furthermore, how the information is collected can represent a problem because direct methods such as cameras invade the personal privacy. Therefore, the main purpose of this research is the development of an affordable and non-intrusive low-cost solution to improve the management of public spaces and ensure the compliance of government regulations regarding the number of occupants inside, implementing unsupervised and semi-supervised machine learning models as well as data fusion from different Internet of Things (IoT) environmental sensor nodes. The preliminary results show that the clusters follow a pattern during the day, therefore this can be used to assign the occupancy levels.

Keywords: Environmental sensors, machine learning, data fusion, occupancy estimation.

1 Introduction

Today, we are facing an unprecedented world health problem derivative from SARS-CoV-2 which has changed all the routines and habits around the world. Preventive measures have been implemented to allow economic and social activities to be carried out while minimizing the impact on people's health. Such measures include the use of face masks, physical distancing, bans on large gatherings, mobility restrictions, among others [4]. It is important to flag out, that even after you're vaccinated, you should keep taking precautions [9].

Therefore, the main purpose of this research is the development of an affordable and non-intrusive low-cost solution to improve the management of public spaces and ensure the compliance of government regulations regarding the

number of occupants inside. The proposed solution monitors air temperature, relative humidity, and barometric pressure and apply unsupervised and semi-supervised machine learning algorithms as well as fusion data techniques to estimate the occupancy levels in enclosed spaces. The proposed device to collect the data is equipped with a micro-controller ESP32 [13], and one BME280 [12] sensor.

This paper is structured as follows. Section 2 discusses the related work. Section 3 presents the Hypothesis and Research Questions. Section 4 shows the research objectives. Section 5 presents the methodology to follow. Finally, the state of the research and conclusions are presented in Section 6 and 7.

2 Related Work

During the past years, several research focused on occupancy detection and estimation had been carried out intending to reduce energy consumption, maximizing comfort, security management, and so on [6, 19, 17].

Solutions for COVID-19 Restrictions

Some researches have proposed solutions to handle the COVID-19 preventive measure. For instance, Longo et al. [7] presents a prototype "Smart Gate", whose main function is to perform people flow monitoring and to keep track of the occupancy levels of both indoor and outdoor spaces. Floris et al. [5] proposes an IoT-based smart building solution for indoor environment management, which aims to provide main functionalities as monitoring environmental parameters of the room, detection of the number of occupants in the room, among others.

Design and Development of an Occupancy Sensor

Researchers from The Tecnologico de Monterrey designed a device which estimating the occupancy levels in closed spaces. The device is equipped with a micro-controller ESP32 [13], a BME280 [12], and other electronic components. The controller is programmed to measure every minute the temperature, barometric pressure, and humidity. The data is sent through Wi-Fi to store it in a data-based cloud. Vela et al. [15] carried out occupancy level estimation research, deploying the proposed device within a university gym and in a living room. The accuracy obtained was around 95.2% to 97% using Support Vector Machine (SVM), K-nearest neighbor (k-NN), and decision tree (DT) models.

3 Hypothesis and Research Questions

It is possible to approximate with an accuracy at least of 90% the occupancy levels in enclosed space using unlabeled fusion data from environmental sensing such as air temperature, relative humidity, and atmospheric pressure. Some *research questions* are:

- How many sensors are necessary to install within a room depending on its dimension?
- How deploy the sensors? (position,height,etc.)

- Is essential to add another environmental variable or non-intrusive sensor to estimate occupancy levels in real-time with an accuracy at least of 90%?
- Is it possible to use unlabeled data for estimating the occupancy levels?
- Could data fusion improve the performance of the unsupervised machine learning model?
- Could externals weather conditions affect the achievement of the accurate estimation? Due to the Covid-19 pandemic in which it is necessary to have natural ventilation (windows and/or door open)

4 Objectives

The general objective of this work is to apply unsupervised and semi-supervised machine learning algorithms as well as data fusion methods to estimate the occupancy levels indoors using only ambient variables. The particular goals be achieve as this research work is conducted are:

- Select three possible scenarios with different design-use and dimensions.
- Set the number of nodes (microcontroller with the sensors) necessaries to deploy inside the place.
- Collect data without labels and data with labels.
- Process the datasets and implement a data fusion method.
- Develop an unsupervised and semi-supervised machine learning algorithm.
- Assess the performance of the models and the number of sensors installed by a square meter.
- Analyze whether natural ventilation through doors and windows affects to the machine learning models performance.

Research Contributions The main contribution of this research will be an unsupervised and semi-supervised system to estimate the occupancy levels indoors, as well as the design of a data fusion framework, and the generation of three robust data sets collected in enclosed spaces which will be available to everyone for future analysis.

5 Methodology

The main activities are described below:

1. ***Selection of the test-bed scenarios:*** Three different places with a specific design use will be selected. For instance, meeting rooms, multi-occupant office, and collaborative spaces.
2. ***Sensor Deployment:*** To set the number of devices deployed within each space will be based on the place's superficial area and the literature review.
3. ***Data collection:*** For 3 weeks, an unlabeled data will be collected.
4. ***Data pre-processing:*** It is essential to seek a missing value and outliers, erroneous measures, remove the noise, and the extraction of features.

5. ***Data Exploration:*** Visualizations such as heat maps, histograms, and 3D scatter plots will be carried out. The main goal is to find patterns and hidden substructure which helps us to interpret the outcomes from the models.
6. ***Data Fusion:*** In this stage, the aim is to archive a better performance of the models, integrating the data collected from multiple sensors.
7. ***Machine Learning model selection:*** Through the literature review, the semi-supervised and the unsupervised algorithms will be selected. The total of clusters and/or classes desired is four that will be identified as empty, low, medium, and high occupancy levels.
8. ***Model evaluation:*** The Internal validation index established the quality of the clustering structure without having access to external information [10]. Nevertheless, the clusters need to be compared with the real occupancy. Evaluating it with metrics as Accuracy, and confusion matrix. Therefore, metrics such as accuracy and confusion matrix will be used.

6 State of the Research

During the year 2021, the research work was focused on improving the proof-of-concept system designed by the peers, it is essential because the next experiments involve the deployment of several devices *in-situ*. Furthermore, preliminary experiments were carried out; first, the k-means algorithm was fitted using the data collected by Vela et al. [15], to verify that *"unlabeled"* data can be used to estimate occupancy levels. Second, the prototype was placed in a multi-occupant office belongs to Tecnologico de Monterrey. The data were collected for two weeks to develop unsupervised algorithms such as k-means, agglomerative, and Fuzzy C-means. Nowadays, the research work will start with the selection of the three test-bed enclosed spaces, the sensor deployment, the collection and analysis of data.

7 Preliminary Results

The preliminary results obtained during February to September, 2021 are presented.

Prototype as a Minimum Viable Product (MVP).

The prototype is currently an MVP. Its components were designed and printed as a printing circuit board (PCB) of 5 cm x 5 cm, covered by an external case (6.5 cm x 6.5 cm x 3 cm). This prototype can be easily installed anywhere in the enclosed space.

K-means for living room data.

From the data, the classes was deleted; then the data were normalized and standardized. Four clusters were set to seek. The results show similarities with the classes from the original data. In other words, the clusters formed can be used as classes (empty, low, mid, high) which indicate the occupancy levels.

K-means, agglomerative and Fuzzy C-means for office data.

The data collected in the multi-occupant office do not have any label, the attendance record was requested to verify the real occupancy; therefore the algorithms had not yet been evaluated as predictors of the occupancy levels. Nevertheless, the evaluation of clusters using Silhouette Index was 60% for k-means, 64% for agglomerative, and 67% for Fuzzy C-means. The cluster obtained follow the patters of the temperature and humidity during the day.

References

1. Aosong Electronics Co. Ltd.: DHT22 (DHT22 also named as AM2302). https://www.sparkfun.com/datasheets/Sensors/Temperature/DHT22.pdf
2. Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W.C., Wang, C.B., Bernardini,S.: The Covid-19 pandemic. Critical Reviews in Clinical Laboratory Sciences 57(6), 365–388 (2020). https://doi.org/10.1080/10408363.2020.1783198, https://doi.org/10.1080/10408363.2020.1783198, pMID: 32645276
3. Fajilla, G., Austin, M.C., Mora, D., Simone, M.D.: Assessment of probabilistic models to estimate the occupancy state in office buildings using indoor parameters and user-related variables. Energy and Buildings 246, 111105 (sep 2021). https://doi.org/10.1016/j.enbuild.2021.111105
4. Fernandez-Carames, T.M.; Froiz-Miguez: An IoT and Blockchain based System for Monitoring and Tracking Real-time Occupancy for COVID-19 Public Safety. In: Proceedings of the 7th International Electronic Conference on Sensors and Applications. Switzerland (Nov 2020). https://doi.org/https://doi.org/10.3390/ecsa-7-08207
5. Floris, A., Porcu, S., Girau, R., Atzori, L.: An IoT-Based Smart Building Solution for Indoor Environment Management and Occupants Prediction. Energies 14(10), 2959 (may 2021). https://doi.org/10.3390/en14102959
6. Han, K., Zhang, J.: Energy-saving building system integration with a smart and low-cost sensing/control network for sustainable and healthy living environments: Demonstration case study. Energy and Buildings 214, 109861 (may 2020). https://doi.org/10.1016/j.enbuild.2020.109861
7. Longo, E., Redondi, A.E.C., Bianchini, M., Bolzan, P., Maffei, S.: Smart gate: a modular system for occupancy and environmental monitoring of spaces. In: 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech). IEEE (sep 2020). https://doi.org/10.23919/splitech49282.2020.9243754
8. Secretaria de Relaciones Exteriores de Mexico: Mexico Covid-19 monitoring system. (2021) https://embamex.sre.gob.mx/eua/index.php/en/2016-04-09-20-40-51/ tourism/1760-mexico-s-covid-19-monitoring-system
9. World Health Organization: Covid-19 advice for the public: Getting vaccinated. (Sep 2021) https://www.who.int/ emergencies/ diseases/ novel-coronavirus -2019/ covid-19-vaccines/ advice
10. Palacio Nino, J.: Evaluation metrics for unsupervised learning algorithms. (2019) https://doi.org/arXiv:1905.05667
11. Patel, A. A.: Hands-On Unsupervised Learning Using Python. Orelly Media,Inc, first edition edn. (Feb 2019)
12. Bosch Sensortec: BME280 Digital humidity, pressure and temperature sensor. Bosch Sensortec, bst-bme280-ds002-15 edn. (Sep 2018), https://www.mouser.com/datasheet/2/783/BST-BME280-DS002-1509607.pdf

13. Espressif Systems: ESP32 Series. (2021) https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_en.pdf
14. Trivedi, D., Badarla, V.: Occupancy detection systems for indoor environments: A survey of approaches and methods. Indoor and Built Environment 29(8), 1053–1069 (2020). https://doi.org/10.1177/1420326X19875621, https://doi.org/10.1177/1420326X19875621
15. Vela, A., Alvarado-Uribe, J., Davila, M., Hernandez-Gress, N., Ceballos, H.G.: Estimating Occupancy Levels in Enclosed Spaces Using Environmental Variables: A Fitness Gym and Living Room as Evaluation Scenarios. Sensors 20(22), 6579 (nov 2020). https://doi.org/10.3390/s20226579
16. Wei, Y., Xia, L., Pan, S., Wu, J., Zhang, X., Han, M., Zhang, W., Xie, J., Li, Q.: Prediction of occupancy level and energy consumption in office building using blind system identification and neural networks. Applied Energy 240, 276–294 (apr 2019). https://doi.org/10.1016/j.apenergy.2019.02.056
17. Yang, Z., Li, N., Becerik-Gerber, B., Orosz, M.: A Multi-Sensor Based Occupancy Estimation Model for Supporting Demand Driven HVAC Operations. In: Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design. SimAUD '12, Society for Computer Simulation International, San Diego, CA, USA (2012)
18. Zheng, A.: Evaluating Machine Learning Models. O'Reilly Media, Inc, first edition edn. (Sep 2015)
19. Zimmermann, L., Weigel, R., Fischer, G.: Fusion of nonintrusive environmental sensors for occupancy detection in smart homes. IEEE Internet of Things Journal 5(4), 2343–2352 (aug 2018). https://doi.org/10.1109/jiot.2017.2752134

ISSN 1870-4069

Clustering Models for the Identification of Coexisting Bacteria in Groups of Patients with the Polymicrobial Syndrome Bacterial Vaginosis (BV)

Henry Jesús Hernández Gómez, Juana Canul Reich

Juarez Autonomous University of Tabasco,
Academic Division of Information Sciences and Technologies,
Mexico

henryhernandezgomez@hotmail.com, juana.canul@ujat.mx

1 Motivation

Bacterial vaginosis is the most frequent infection in women of child-bearing age [2]. The pathogenesis is unclear, but involves overgrowth due to multiple bacterial pathogens and a decrease in the usual vaginal flora with a predominance of lactobacillus [7, 5]. Its complications, such as endometritis, salpingitis, oophoritis [4], preterm premature rupture of membranes (PPROM), and chorioamnionitis [6]. Particularly, it is well known that a BV-positive case is a consequence of a disequilibrium state of bacteria; it is not a unique bacterium but a coexistence of bacteria leading to a BV-positive condition. However, these bacteria may differ from patient to patient.

Thus, the present research arises from the need to support physicians in understanding the unclear pathogenesis of BV by identifying clusters of patients that presumably share coexisting bacteria in a grouped form, and from there, to analyze the contexts of bacterial coexistence. So it is our interest to tackle this problem using a clustering approach.

The dataset used for this research was generated by the laboratory of research in infectious and metabolic diseases of the Juarez Autonomous University of Tabasco. It was obtained as part of research in molecular epidemiology in BV during the years 2016 to 2018 in an urban population in southeastern Mexico [8]. The dataset consists of 201 patient records with 19 attributes, which was explored through AHC methods without performing any pre-processing process.

This study contributes to the effort to provide information that helps to learn more about the coexisting bacteria in the groups of patients diagnosed as BV-positive. Furthermore, the benefit of identifying the groups translates into the selection of specific treatments according to the bacteria coexisting in each group. It also becomes a support tool to obtain a priori knowledge of the contexts that may occur in clinical cases.

2 Previous Works in the Area

Nowadays, there exist few research studies that have conducted an analysis of the BV disease using ML methods.

Song et al. [9] contributed to integrate Superpixel methods with Deep Learning methods based on convolutional neural network (CNN) for the automatic assisted diagnosis of BV. The classification results yield an accuracy of 99%, the sensitivity of 100% and specificity of 98.04%.

Baker et al. [1] built a classification model by breaking down the groups of microbes based on their correlation. Likewise, it reduced the number of factors, increasing the interpretability of the classification models. The classifications were made using Genetic Programming, Random Forest, and Logistic Regression, the precision of the models was evaluated using ROC curves. The accuracy obtained from the models was between 90% and 95% when they were classified using the dataset with the Nugent score.

Cruciani et al. [3] designed a new phylogenetic microarray-based tool (VaginArray) that includes 17 probe sets specific for the most representative bacterial groups of the human vaginal ecosystem. The VaginArray was applied to evaluate the efficacy of rifaximin vaginal tablets for the treatment of BV. The results showed the ability of rifaximin to reduce the growth of various BV-related bacteria (Atopobium vaginae, Prevotella, Megasphaera, Mobiluncus, and Sneathia spp.)

3 Research Objectives

To build a bacterial community clustering model of the polymicrobial syndrome bacterial vaginosis (BV) that allows identifying characteristics of similarity and dissimilarity of the clusters, as well as bacterial coexistence contexts with biological significance.

4 Methodology

In order to achieve the stated research objective, it was necessary to construct a study scheme that shows the clustering approaches considered and the phases to be performed that will lead to visual clustering models that allow the analysis of bacterial coexistence contexts, as shown in Figure 1.

5 State of the Research

Currently, the research shows the progress of more than 75% of the activities carried out. The activities involved in the construction and validation of results through metrics were completed for two of the three investigated approaches.

Furthermore, agglomerative hierarchical clustering and partitional clustering have been constructed. Finally, each model was converted to a data visualization

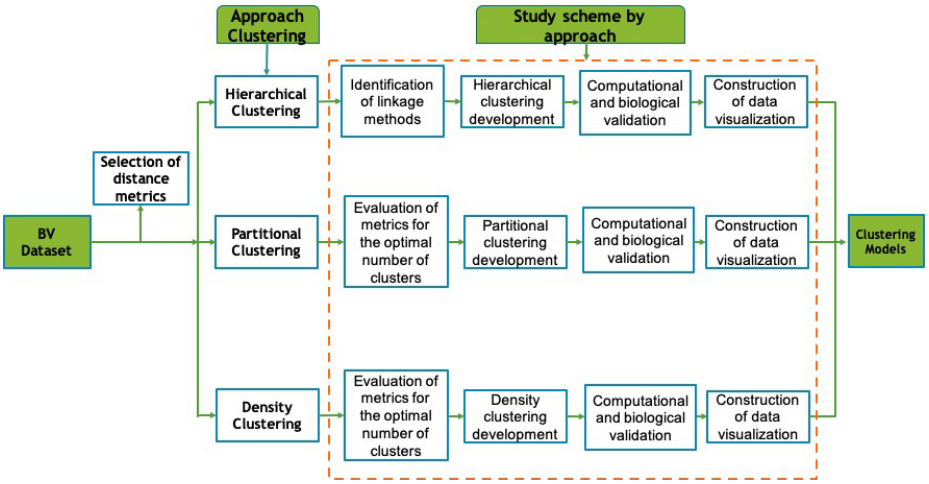

Fig. 1. Study scheme.

that allows analysis of the contexts of bacterial coexistence between groups of patients. The density clustering algorithms so far are about to start the process of constructing the data visualizations.

As the research is pioneering the use of clustering methods to create visual models for further analysis of bacterial coexistence contexts, each model is subjected to biological validation by an expert in the field who evaluates the biological significance of the groups.

6 Preliminary Results

The overall results that have been achieved so far in BV research through clustering models are:

Two experiments were designed and evaluated in the agglomerative hierarchical clustering (AHC) methods through five linkage methods and a distance measure. The first experiment aimed to identify clusters with the presence and absence of VB. The result obtained was a segmentation of the dataset into two groups, one with elements with positive classes and a second group with elements of negative and indeterminate classes.

The clustering partitioning (PC) methods were evaluated through four methods from the same experiments perspective as the AHCs. The findings of the PCs assessed in the first experiment show a similarity of clustering to the AHCs. However, in the second experiment, the algorithms show a dissimilar clustering with the common VB-positive diagnostic. From the findings across the two clustering approaches evaluated, we can infer that the best performance achieved for the second experiment is from the AHCs.

We have created data visualization tools for further analysis of bacterial coexistence. The identified clusters show a context of bacterial coexistence of

two anaerobic bacteria with a prevalence of more than 30%. The findings of the dissimilar clusters with a common diagnosis, there is a coexistence of three bacteria with a prevalence of more than 30%.

We have also experimented with other measures of similarity to improve the results of experiment two of the PC.

References

1. Beck, D., Foster, J.A.: Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PloS one 9(2), e87830 (2014)
2. Coleman, J.S., Gaydos, C.A.: Molecular diagnosis of bacterial vaginosis: an update. Journal of clinical microbiology 56(9), e00342–18 (2018)
3. Cruciani, F., Biagi, E., Severgnini, M., Consolandi, C., Calanni, F., Donders, G., Brigidi, P., Vitali, B.: Development of a microarray-based tool to characterize vaginal bacterial fluctuations and application to a novel antibiotic treatment for bacterial vaginosis. Antimicrobial agents and chemotherapy 59(5), 2825–2834 (2015)
4. for Disease Control, C., Prevention, et al.: Sexually transmitted diseases treatment guidelines. Morbid Mortal 42, 57–59 (1993)
5. Muzny, C.A., Taylor, C.M., Swords, W.E., Tamhane, A., Chattopadhyay, D., Cerca, N., Schwebke, J.R.: An updated conceptual model on the pathogenesis of bacterial vaginosis. The Journal of infectious diseases 220(9), 1399–1405 (2019)
6. Prajarto, H.W., Pramono, B.A.: The association of cervical length, bacterial vaginosis, urinary tract infection and premature rupture of membranes to the imminent preterm labour. Diponegoro International Medical Journal (DIMJ) 1(2), 10–6 (2020)
7. Reid, G.: Is bacterial vaginosis a disease? Applied microbiology and biotechnology 102(2), 553–558 (2018)
8. Sanchez Garcia, E.K., Contreras Paredes, A., Martinez Abundis, E., Garcia Chan, D., Lizano, M., de la cruz Hernandez, E.: Molecular epidemiology of bacterial vaginosis and its association with genital microorganisms in asymptomatic women. Journal of medical microbiology 68(9), 1373–1382 (2019)
9. Song, Y., Ni, D., Zeng, Z., He, L., Chen, S., Lei, B., Wang, T.: Automatic vaginal bacteria segmentation and classification based on superpixel and deep learning. Journal of Medical Imaging and Health Informatics 4(5), 781–786 (2014)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación