

Hybridation d'un agent conversationnel avec des plongements lexicaux pour la formation au diagnostic médical

Fréjus A. A. Laleye¹ Antonia Blanié² Antoine Brouquet² Dan Benhamou²
Gaël de Chalendar¹

(1) CEA, LIST, Laboratoire d'Analyse Sémantique Texte et Image, Gif-sur-Yvette, F-91191

(2) Laboratoire de Formation par la Simulation et l'Image en Médecine et en Santé (LabForSIMS), Faculté de Médecine Paris-Sud, 94275 Le Kremlin Bicêtre

frejus.laleye@cea.fr, gael.de-chalendar@cea.fr, antonia.blanie@aphp.fr,
antoine.brouquet@aphp.fr, dan.behnamou@aphp.fr

RÉSUMÉ

Dans le contexte médical, un patient ou médecin virtuel dialoguant permet de former les apprenants au diagnostic médical via la simulation de manière autonome. Dans ce travail, nous avons exploité les propriétés sémantiques capturées par les représentations distribuées de mots pour la recherche de questions similaires dans le système de dialogues d'un agent conversationnel médical. Deux systèmes de dialogues ont été créés et évalués sur des jeux de données collectées lors des tests avec les apprenants. Le premier système fondé sur la correspondance de règles de dialogue créées à la main présente une performance globale de 92% comme taux de réponses cohérentes sur le cas clinique étudié tandis que le second système qui combine les règles de dialogue et la similarité sémantique réalise une performance de 97% de réponses cohérentes en réduisant de 7% les erreurs de compréhension par rapport au système de correspondance de règles.

ABSTRACT

Hybridization of a conversational agent with word embeddings for medical diagnostic training

In the medical domain, virtual patient or doctor dialogue systems help to train students autonomously to medical diagnosis via simulation. In this work, we exploited the semantic properties captured by distributed word representations to search for similar questions in the dialogue system of a medical conversational agent. We created two dialogue systems that were evaluated on datasets collected during tests with students. The first system based on hand-crafted rules obtains 92% of correct responses on the studied clinical case while the second system that combines rules and semantic similarity achieves a score of 97%. It represents an error reduction of 7% as compared to the rules-only-based system.

MOTS-CLÉS : Agents conversationnels, Chatbots, Médecine, Formation, Similarité sémantique, Plongements Lexicaux.

KEYWORDS: Conversational Agents, Chatbots, Medical Education, Semantic Similarity, Word Embeddings.

1 Introduction

La pratique du diagnostic médical s'apprend traditionnellement "au lit du malade". Les cours théoriques sont complétés par des stages dans les services hospitaliers. L'étudiant en médecine y observe la pratique des médecins et des internes et pratique lui-même sous le contrôle de ceux-ci. Ce type de formation présente le désavantage de confronter d'emblée l'étudiant en médecine à des situations complexes sans apprentissage pratique (technique et humain) préalable. Il doit gérer en même temps les relations avec des personnes en souffrance et la mobilisation de connaissances complexes et encore incomplètes.

Il apparaît donc utile de pouvoir s'entraîner avant de se confronter à ses premiers patients. Mais pour que cela soit réaliste, il ne peut pas se faire avec des pairs qui joueraient le rôle de patients. Le réalisme serait insuffisant. Il ne peut pas non plus travailler avec des acteurs, le coût serait trop élevé.

Les progrès des outils de réalité virtuelle permettent de plonger à moindre coût l'étudiant dans un environnement réaliste et maîtrisé pédagogiquement. Les chatbots standards permettent de gérer les bases d'un dialogue standardisé entre l'étudiant et le patient virtuel mais sont insuffisants dans le sens ou le moindre écart par rapport au scénario où au lexique prévus les mettent en échec.

Dans cet article, nous montrons comment exploiter des techniques de similarité sémantique fondées sur l'utilisation de plongements lexicaux pour produire un agent conversationnel acceptable pour l'apprentissage des étudiants en médecine.

2 Etat de l'art

Les récents progrès des technologies de reconnaissance vocale, du traitement du langage naturel et de l'intelligence artificielle ont conduit à l'émergence des agents conversationnels dans différents domaines de la vie tels que la santé, les finances ou l'éducation. Dans le domaine de la santé, on assiste à l'expansion des patients ou médecins virtuels utilisés pour interagir avec l'homme dans des simulations de scénarios cliniques à des fins de formation, d'éducation ou d'évaluation médicale. Pour la formation en médecine, les apprenants jouent le rôle d'un soignant pour diagnostiquer le patient virtuel et prescrire des traitements avec des scénarios d'entrevue cliniques et pédagogiques validés (Saffari *et al.*, 2014).

Les patients virtuels prennent différentes formes selon les objectifs pédagogiques visés. On les retrouve dans des simulateurs de réalité virtuelle pour la formation en endoscopie (Harpham-Lockyer *et al.*, 2015). Ils sont aussi utilisés, pour enseigner les techniques d'examen oral aux internes en médecine d'urgence (McGrath *et al.*, 2015), pour permettre aux infirmières de développer des compétences en soins aigus telles que l'évaluation et la gestion de la détérioration clinique (Liaw, 2015) et aux stagiaires en médecine de pratiquer les techniques de raisonnement clinique (Close *et al.*, 2015; Kleinert, 2015). Plusieurs programmes en médecine ont déjà intégré les simulations incluant des mannequins pour évaluer la compétence et la confiance des apprenants (Taglieri *et al.*, 2017). Il ressort des différentes évaluations que l'utilisation des patients virtuels fournit des pratiques supplémentaires aux apprenants en dehors des travaux pratiques classiques et améliore leurs performances sur les cas réels. (Smith & Waite, 2017) ont montré que la technologie des patients virtuels peut améliorer les performances des apprenants sur les questions de consultations cliniques.

De nombreuses recherches sont en cours pour améliorer l'efficacité des agents conversationnels

virtuels afin de répondre aux besoins sans cesse croissants des professionnels de la santé. Tous les projets en cours ont pour objectif d'augmenter et d'améliorer l'interaction entre les patients et les médecins (Bioulac *et al.*, 2018; Campillos *et al.*, 2016, 2017). Cette interaction, dans les simulations employant les agents conversationnels dialoguants, se caractérise par : (i) le type de technologie (plateforme hébergeant l'agent conversationnel) à savoir les smartphones (Miner *et al.*, 2016; Ireland *et al.*, 2016), les ordinateurs portables ou de bureau (Tanaka *et al.*, 2017; Philip *et al.*, 2017) et les plateformes multimodales (Lucas *et al.*, 2017); (ii) la stratégie de gestion du dialogue à savoir la stratégie à états finis où le dialogue est une séquence d'étapes prédéfinies (Tanaka *et al.*, 2017; Philip *et al.*, 2017; Lucas *et al.*, 2017); la stratégie fondée sur le contenu où le flux de dialogue n'est pas prédéfini mais dépend du contenu fourni par l'utilisateur (Ireland *et al.*, 2016); la stratégie fondée sur l'agent où le dialogue s'effectue entre deux agents capables de raisonner (Miner *et al.*, 2016); (iii) l'initiative dans le dialogue où c'est soit l'utilisateur qui initie la conversation (Miner *et al.*, 2016), soit l'agent virtuel qui conduit la conversation (Tanaka *et al.*, 2017; Philip *et al.*, 2017; Lucas *et al.*, 2017), soit les deux qui peuvent mener la conversation (Fitzpatrick *et al.*, 2017; Ireland *et al.*, 2016).

Dans notre système, les agents conversationnels sont intégrés dans une simulation en Réalité Virtuelle. Ils marient états finis et conversation libre. Certains agents mènent la conversation tandis que d'autres suivent la conversation menée par l'apprenant.

(Jin *et al.*, 2017), réalisant la même tâche de compréhension de questions par un patient virtuel, ont associé un réseau de neurones convolutif et le système ChatScript par un système de combinaison utilisant un classifieur binaire sur des questions collectées lors des séances avec des étudiants en médecine.

3 Architecture du système

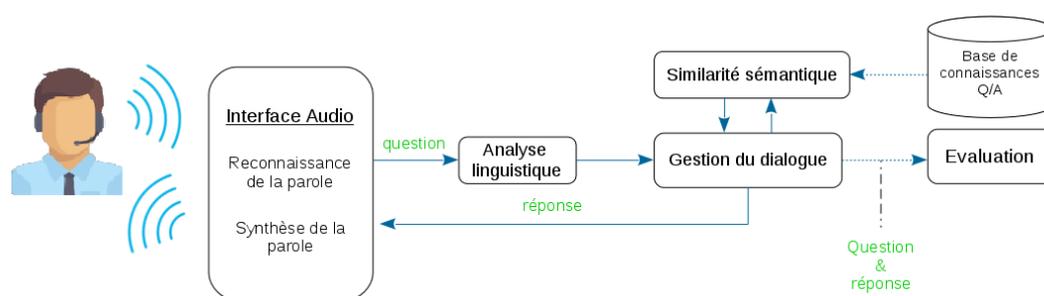


FIGURE 1 – Architecture de l'interaction Apprenant-Agent

La figure 1 présente une vue globale de l'interaction entre l'apprenant et l'agent conversationnel. Le système proposé est conçu pour la consultation d'une urgence chirurgicale abdominale et intègre à la fois la patiente et les médecins (radiologue, chirurgien, etc.) virtuels. La stratégie et l'initiative du dialogue dépendent du profil de l'agent conversationnel. Le système est fondé à la fois sur une stratégie de gestion du dialogue à états finis pour les agents virtuels médecins et une stratégie fondée sur le contenu fourni par l'apprenant pour l'agent virtuel patient. L'architecture présente cinq modules :

- l'interface audio : l'interaction avec l'apprenant se fait via la voix ; l'interface audio constitue le module d'entrée et de sortie du système ; elle intègre un moteur de reconnaissance automatique de la parole (Google Cloud Speech) et un ensemble de générateurs de synthèse vocale (MaryTTS et SVOX Pico) ;

- l’analyse linguistique : l’analyseur linguistique, fondé sur LIMA (Besançon *et al.*, 2010), se charge de filtrer le texte en supprimant les mots vides et de normaliser le texte renvoyé par le moteur de reconnaissance de la parole ;
- la gestion du dialogue : le gestionnaire de dialogue, fondé sur ChatScript (Wilcox & Wilcox, 2013), est un système à base de règles dont les fonctions sont l’interprétation des questions à partir de correspondances de modèles, la génération de réponses et la gestion des modules de recherche de similarité sémantique et d’évaluation. Il se charge, (i) ; de retrouver la règle correspondante à la question posée. Les règles sont des sortes d’expression régulières dont les atomes sont les mots. Des classes d’équivalence de mots définissent des concepts. Les règles peuvent prendre en compte un contexte de quelques questions mais le système n’intègre pas de véritable gestionnaire de dialogue. Nous enrichissons ce fonctionnement en permettant au systèmes d’accéder au informations déjà recueillies à propos du patient ou de la patiente. (ii) Si la règle est trouvée, ChatScript produit une réponse prévue à l’avance. (iii) Sinon la question est envoyée au module de recherche de similarité sémantique qui lui renvoie, à partir de la base de connaissances, une question sémantiquement proche. (iv) Il recherche alors à nouveau la règle correspondant à cette question similaire et génère de même la réponse ;
- la recherche de similarité sémantique : l’objectif de ce module est de réduire les limites de compréhension du système à base de règles en assurant une large couverture des questions ; le module de similarité sémantique, fondé sur les plongements lexicaux, calcule une mesure de similarité entre deux questions q_i et q_j (avec q_i la question de l’apprenant et q_j une question déjà associée ou non à des règles de dialogue). Il est composé d’un classifieur neuronal de textes qui, avec comme entrées les plongements lexicaux pré-entraînés, classe une question q_i dans une catégorie c (champ de données cliniques) et d’un calculateur de distances cosinus (voir Eq. 1) entre les vecteurs de plongements lexicaux de la question q_i et l’ensemble des questions de la catégorie c . Le résultat est la question de c la plus sémantiquement proche de q_i en terme de distance. La distance est calculée à partir de la somme pondérée des vecteurs mots (w_{i1}, \dots, w_{in}) de q_i et $(w'_{j1}, \dots, w'_{jn})$ de q_j et se présente comme suit :

$$dist(q_i q_j) = 1 - \cos \left(\sum_{k=0}^N \sigma_k v(w_{ik}); \sum_{k'=0}^{N'} \sigma'_{k'} v(w'_{jk'}) \right) \quad (1)$$

où σ_k et $\sigma'_{k'}$ représentent respectivement la pondération des mots k et k' avec leurs fréquences inverses en documents.

- l’évaluation : ce module met à la disposition des formateurs des informations issues de la simulation pour une évaluation pédagogique du raisonnement clinique de l’apprenant·e.

4 Évaluation et résultats

Dans ce travail, nous présentons les résultats issus de l’évaluation de l’agent conversationnel *patient*. Il s’agit de l’interaction entre l’étudiant et la patiente virtuelle pour une stratégie diagnostique d’une urgence chirurgicale abdominale. L’objectif est d’évaluer la capacité de l’agent conversationnel à fournir des réponses cohérentes aux questions posées par l’étudiant·e.

4.1 Description des données

Nous sommes partis des types de questions ("*symptôme : où avez-vous mal ?*", "*traitement : est-ce que vous êtes sous traitement ?*", "*antécédent : avez-vous des antécédents médicaux*") pour écrire les règles ChatScript des différents agents conversationnels. Pour entraîner le classifieur neuronal de textes et le calculateur de distances sémantiques, nous avons collecté les sous-titres de dialogues en français de la série télévisée *ChicagoMed* et les données utilisées par (Campillos *et al.*, 2017) pour la classification automatique de questions médecin-patient. Les questions collectées ont été manuellement annotées suivant les catégories ci-dessus de questions d'une consultation chirurgicale pour des douleurs abdominales. Elles sont utilisées pour construire la base de connaissances du module de similarité sémantique.

Des réponses ont été manuellement associées aux questions afin de construire des jeux de questions/réponses pour l'évaluation globale de l'approche à base de règles et l'approche combinant les règles et la similarité sémantique fondée sur les plongements lexicaux. A la suite des séances de tests effectuées, nous avons recueilli un jeu de données comportant 362 questions. Aussi, nous avons extrait des sous-titres de dialogues de la série *ChicagoMed* un ensemble de 202 questions contenant à la fois des questions d'une consultation pour des douleurs abdominales et des questions qui s'éloignent de ce cas clinique. Le but de ce jeu de données est d'évaluer la capacité de l'agent conversationnel à couvrir un maximum de questions avec le même nombre de règles de dialogue. Chaque jeu de données intègre différentes formulations des mêmes questions. Le tableau 1 présente quelques exemples de question et de leurs reformulations.

Type	Questions	Reformulations	Réponses
Symptôme	Est-ce que vous avez frissonner ?	Avez-vous eu froid ?	Oui j'ai eu froid
Sujet de consultation	Qu'est-ce qui vous amène ici ?	Qu'est-ce qui ne va pas madame ?	J'ai très mal au ventre
Symptôme	Où se situe la douleur ?	où est-ce que la douleur est localisée ?	J'ai mal au niveau du ventre
Personnel	Que faites-vous dans la vie ?	Quel est votre métier	Je suis enseignante

TABLE 1 – Exemples de reformulations de questions.

4.2 Métriques d'évaluation

L'évaluation est réalisée sur la base des types de réponses des deux systèmes. Les métriques pour définir quantitativement les performances de chaque système incluent :

- le nombre de réponses cohérentes équivalant à une bonne compréhension des questions ;
- le nombre de réponses incohérentes équivalant à une mauvaise compréhension des questions ;
- le nombre de réponses "*je n'ai pas compris*" équivalant à une non compréhension des questions.

La cohérence et l'incohérence des réponses ont été jugées par deux médecins experts auteurs du présent article.

4.3 Résultats et analyses

La performance de chacun des systèmes est mesurée sur l'ensemble des questions collectées et annotées. Pour les 15 étudiants testeurs, nous avons obtenu un ensemble de 362 questions qui a servi de données pour l'évaluation.

Dans un premier temps, nous avons évalué la capacité des classifieurs (intégrés au module de similarité sémantique) à identifier la catégorie d'une question. Un jeu de 90k questions/réponses a été extrait de l'ensemble des données collectées pour servir de données d'apprentissage et de validation.

Chaque catégorie est traitée comme un problème de classification binaire. Pour l'évaluation, nous avons utilisé une validation croisée en 10 parties. Pour chaque partie, nous avons divisé les données en deux ensembles (apprentissage et validation) avec un ratio de 90/10. La précision obtenue après apprentissage et évaluation de chaque classifieur est présentée à la figure 2.

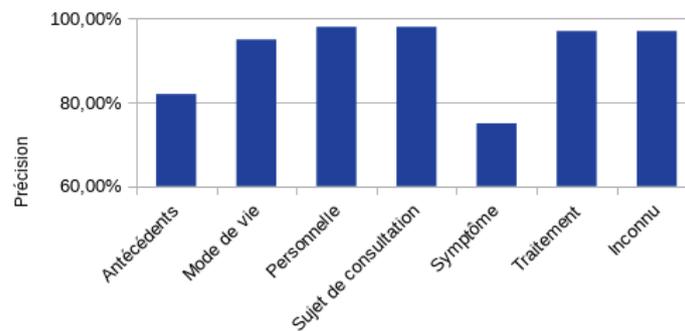


FIGURE 2 – Précision de chaque classifieur de type de questions après la validation croisée.

Nous avons construit le calculateur de distances sémantiques à partir des représentations distribuées des phrases issues des données collectées. Les représentations distribuées des phrases ont été entraînées de façon non supervisée sur un jeu de 45 000 phrases. Le classifieur et le calculateur ont été combinés au système de gestion à base de règles pour une comparaison de performances.

Questions	Réponses	Type de réponses
Est-ce que vous avez frissonner	Oui j'ai froid	réponse cohérente
Quelle est la couleur de vos urines	Je ne sais pas docteur, j'avais fini de manger	réponse incohérente
La pizza était au frigidaire	Je n'ai pas compris.	je n'ai pas compris

TABLE 2 – Exemple de réponses fournies par le système combiné.

Le tableau 2 présente un exemple de chaque type de réponse. Le tableau 3 montre leurs scores. Pour le corpus issu des tests effectués avec les étudiants, le système combiné réalise une performance de 97% de questions bien comprises (*réponses cohérentes*) en améliorant d'environ 5 points le score du système à base de règles. En évaluant la capacité des deux systèmes à comprendre une question et à fournir une réponse cohérente ou non, le système combiné a réduit de 7% à presque 0% le taux de réponses *je n'ai pas compris*. Des réponses, parfois incohérentes, sont extraites de la base des questions/réponses du module de similarité lorsqu'il n'y a pas de règles trouvées. Ceci justifie l'augmentation à (3%) du taux de réponses incohérentes. Cette performance vient réduire la frustration de l'étudiant, souvent générée par les réponses *je n'ai pas compris* de l'agent conversationnel lors d'une simulation, qui impacte sur le bon déroulement des interactions dans le sens où l'apprenant doute de la formulation de sa question ou des connaissances de la patiente virtuelle.

Corpus	Système fondé sur les règles			Système combiné (règles + similarité)		
	<i>réponses co-hérentes</i>	<i>réponses in-cohérentes</i>	<i>je n'ai pas compris</i>	<i>réponses co-hérentes</i>	<i>réponses in-cohérentes</i>	<i>je n'ai pas compris</i>
Tests	92%	1%	7%	97%	3%	0%
Sous-titres	65%	2%	33%	94%	5%	1%

TABLE 3 – Évaluation du système fondé sur les règles et de sa combinaison avec la similarité sémantique.

Le taux 33% de questions non comprises par le système à base de règles démontre l'écart entre les questions du corpus issu des sous-titres et le cas clinique étudié. Il montre que les règles de dialogue manuelles éprouvent des difficultés à traiter des questions qui s'éloignent du contexte. L'ajout de la mesure sémantique a réduit de 33% à 1% le taux de réponses *je n'ai pas compris* en augmentant jusqu'à 94% le taux de réponses cohérentes. Ceci montre que, sans ajout de nouvelles règles, la similarité sémantique complète efficacement le système fondé sur les règles de dialogue, rendant le système combiné plus performant.

La combinaison des réseaux de neurones et de ChatScript réalisée par (Jin *et al.*, 2017) a permis d'obtenir une précision de 89,3% en augmentant de plus des deux tiers la précision du réseau de neurones et en réduisant de 47% les erreurs sur ChatScript.

5 Conclusion

Nous avons décrit dans cet article un système de dialogue à interface vocale pour la formation des étudiants en médecine au diagnostic des urgences chirurgicales. Notre système marie les capacités de description d'un scénario de dialogue par des règles à la résilience fournie par la similarité sémantique fondée sur des plongements lexicaux. Nous avons atteint un taux de compréhension qui rend le système utilisable par les étudiants. Dans la suite, nous allons concevoir avec les médecins participant au projet les protocoles permettant d'évaluer l'apport pour la formation des étudiants du système global intégrant agent conversationnel et environnement virtuel dynamique.

Remerciements

Nous remercions Leonardo Campillos Llanos et ses collègues du LIMSI pour nous avoir donné accès à leur corpus d'interactions entre médecins et agent conversationnel. Merci aussi aux étudiants en médecine de l'Université Paris-Sud qui ont bien voulu participer aux tests de notre propre agent conversationnel.

Références

- BESANÇON R., DE CHALENDAR G., FERRET O., GARA F., LAÏB M., MESNARD O. & SEMMAR N. (2010). Lima : A multilingual framework for linguistic analysis and linguistic resources development and evaluation. In *LREC*, p. 3697–3704, Malte.
- BIOULAC S., DE SEVIN E., SAGASPE P., CLARET A., PHILIP P., MICOULAUD-FRANCHI J. & BOUVARD M. (2018). Qu'apportent les outils de réalité virtuelle en psychiatrie de l'enfant et l'adolescent? *L'Encéphale*, **44**(3), 280 – 285.
- CAMPILLOS L., BOUAMOR D., BILINSKI E., LIGOZAT A.-L., ZWEIGENBAUM P. & ROSSET S. (2016). Integrating a dialogue system into a virtual patient consultation. In *AMIA Annual Fall Symposium*, Chicago, USA.
- CAMPILLOS L., ROSSET S. & ZWEIGENBAUM P. (2017). Automatic classification of doctor-patient questions for a virtual patient record query task. In *16th BioNLP 2017 Workshop*, p. 333–341, Vancouver, Canada.
- CLOSE A., GOLDBERG A., HELENOWSKI I., SCHULLER M., DAROSA D. & FRYER J. (2015). Beta test of web-based virtual patient decision-making exercises for residents demonstrates discriminant validity and learning. *Journal of Surgical Education*, **72**(6), 130–136.
- FITZPATRICK K. K., DARCY A. & VIERHILE M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot) : A randomized controlled trial. *JMIR Mental Health*, **4**, 19.
- HARPHAM-LOCKYER L., LASKARATOS F.-M., BERLINGIERI P. & EPSTEIN O. (2015). Role of virtual reality simulation in endoscopy training. *World Journal of Gastrointestinal Endoscopy*, **7**(18), 1287–1294.
- IRELAND D., ATAY C., LIDDLE J., BRADFORD D., LEE H., RUSHIN O., MULLINS T., ANGUS D., WILES J., MCBRIDE S. & VOGEL A. (2016). Hello harlie : Enabling speech monitoring through chat-bot conversations. *Studies in health technology and informatics*, **227**, 55–60.
- JIN L., WHITE M., JAFFE E., ZIMMERMAN L. & DANFORTH D. (2017). Combining cnns and pattern matching for question interpretation in a virtual patient dialogue system. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 11–21, Copenhagen, Denmark.
- KLEINERT E. A. (2015). Web-based immersive virtual patient simulators : Positive effect on clinical reasoning in medical education. *Journal of Medical Internet Research*, **17**(11), 263.
- LIAW E. A. (2015). Designing and evaluating an interactive multimedia web-based simulation for developing nurses' competencies in acute nursing care : randomized controlled trial. *Journal of Medical Internet Research*, **17**(1).
- LUCAS G. M., RIZZO A., GRATCH J., SCHERER S., STRATOU G., BOBERG J. & MORENCY L.-P. (2017). Reporting mental health symptoms : Breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, **4**.
- MCGRATH J., NICHOLAS K., DOUGLAS D., DAVID P. B., SORABH K., DANIEL R. M., ROLLIN N., NICOLE, V. DAVID P. W. & RICHARD N. (2015). Virtual alternative to the oral examination for emergency medicine residents. *The Western Journal of Emergency Medicine*, **16**(2), 336–343.
- MINER A., MILSTEIN A., SCHUELLER S., HEGDE R., MANGURIAN C. & LINOS E. (2016). Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA Internal Medicine*, **176**(5), 619–625.

- PHILIP P., MICOULAUD-FRANCHI J.-A., SAGASPE P., DE SEVIN E., OLIVE J., BIOULAC & SAUTERAUD A. (2017). Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. In *Scientific reports*, volume 7, p. 1–7.
- SAFFARI Z., TAKMIL F. & ARABZADEH R. (2014). The role of educational technology in medical education. *Journal of Advances in Medical Education & Professionalism*, **2**(4), 183.
- SMITH M. & WAITE L. (2017). Utilization of a virtual patient for advanced assessment of student performance in pain management. *Journal of Currents in Pharmacy Teaching and Learning*, **9**(5), 893–897.
- TAGLIERI C. A., CROSBY S. J., ZIMMERMAN K., SCHNEIDER T. & PATEL D. K. (2017). Evaluation of the use of a virtual patient on student competence and confidence in performing simulated clinic visits. *American Journal of Pharmaceutical Education*, **81**(5), 87.
- TANAKA H., NEGORO H., IWASAKA H. & NAKAMURA S. (2017). Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLOS ONE*, **12**(8), 1–15.
- WILCOX B. & WILCOX S. (2013). Making it real : Loebner-winning chatbot design. In *Arbor*, volume 189, p. a086.

