A Framework to support the Experimental Evaluation Process of the Pedagogical Conversational Systems*

Leo Natan Paschoal

Institute of Mathematics and Computer Sciences – University of São Paulo
paschoalln@usp.br

Abstract. Conversational systems are applications whose main characteristic is interaction with humans through spoken or written natural language. Initially, they have been established for educational purposes, dedicated to supporting students while learning new subjects, motivating them while carrying out activities, solving their questions, indicating new learning materials, among others. Several studies show the benefits of using conversational systems how a support tool, promoting the application of active learning methodologies. However, recent studies have mentioned that these systems are not being adequately evaluated since it is not possible to observe systematization in the evaluations, the selection of variables is not standardized, and there is no replicated study yet. Experimental Software Engineering presents robust procedures for experimentation. It can indicate better strategies to evaluate this type of system and offer subsidies to improve the visibility of the potential of these systems as a learning support mechanism. In this project, we intend to collaborate whit the evaluation process of the conversational systems, proposing a framework to support the planning, execution, and reporting of the experimental studies. The area lacks adequate methodologies for reliable experimental studies and, therefore, this project can contribute to improving the quality of conversational systems.

Keywords: Chatbots · Experimentation · Dialogue Systems.

1 Introdução

Sistemas conversacionais representam aplicações que são capazes de processar e emitir a linguagem humana, por meio de texto ou voz [1]. Exemplos comuns dessas aplicações são os assistentes virtuais *Siri, Google Assistant e Cortana* [12]. Além de serem reconhecidos como assistentes virtuais, na literatura acadêmica, é possível identificá-los como agentes conversacionais, *chatbots*, *chatterbots*, sistemas de diálogo, dentre outros [13]. Devido às habilidades do processamento de linguagem natural, essas aplicações têm sido estabelecidas para contextos

* This work is funding by the University of São Paulo and by CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil - Finance Code 001.

de uso específicos, em diferentes áreas do conhecimento humano. Também são definidos para serem usados em conjunto com outros sistemas. Um exemplo é sua integração com sistemas de informação do tipo CRM (*Customer Relationship Management*), para prestar atendimento aos usuários [24].

Sem dúvida, o maior interesse para o uso desses sistemas é no domínio da educação. Isso pode ser observado em recentes estudos secundários que buscam definir um panorama sobre os sistemas conversacionais produzidos para fins educacionais (*e.g.*, [7,8]). Por intermédio dessas pesquisas, é possível observar que os sistemas conversacionais educacionais/pedagógicos estão sendo produzidos para apoiar temáticas específicas, como o ensino de física, matemática, ciência da computação, engenharias e saúde. Além disso, eles são definidos com propósitos diferentes, na medida em que podem ser usados para solucionar dúvidas de alunos que não compreenderam um assunto, repassar informações que sejam relevantes, fornecer *feedback* para as ações realizadas pelos estudantes, usar discurso motivacional para estimular os estudantes, contribuir com a aquisição de conhecimento do aluno, contribuir com o desenvolvimento de competências sociais, dentre outros. Ainda, eles podem ser definidos para se comportarem como tutores, professores ou estudantes.

A exploração do uso de sistemas conversacionais pedagógicos em cursos de plataformas MOOCs (*Massive Open Online Courses*) tem se intensificado [2], uma vez que esses cursos possuem uma ampla quantidade e diversidade de estudantes, e nem sempre há um professor disponível para atender aos questionamentos elaborados pelos alunos. Em trabalhos anteriores, o autor explorou o uso desses sistemas como apoio a práticas de ensino baseadas em metodologias ativas, como *Flipped Classroom* [16]. Apesar de serem estimulados para uso em práticas ativas, esses sistemas também podem ser explorados no ensino tradicional (presencial ou a distância) [11]. Nesse sentido, podem ser usados em conjunto com outros recursos educacionais, tais como ambientes virtuais de aprendizagem, redes sociais, mundos virtuais e jogos educacionais.

Dado a versatilidade e aplicabilidade desses sistemas, era esperado que existissem mais estudos na literatura sobre a implantação dos mesmos em práticas educacionais. Todavia, observa-se que a comunidade não tem apresentado relatos sobre a implantação efetiva desses sistemas em cenários de ensino, pois a maioria dos trabalhos tem focado em descrever o desenvolvimento do sistema conversacional pedagógico e tem se distanciado em explorar sua implantação. Foram identificados estudos dedicados a avaliar, por exemplo, a usabilidade de sistemas conversacionais pedagógicos [15]. Também foram encontrados estudos que descrevem o *feedback* dos alunos ou do professor sobre o benefício do agente conversacional [8,22]. No entanto, esses estudos também são feitos pelos autores para comprovar alguma funcionalidade ou conferir se ele atende a algum objetivo específico [7]. Portanto, relatos de experiência da implantação dos sistemas conversacionais pedagógicos não são comuns em fóruns de divulgação científica ou em artigos científicos.

Devido a falta de relatórios ou artigos científicos sobre as experiências em implantações desses sistemas, não existem lições a serem aprendidas. Por

outro lado, isso pode levar a alguns questionamentos associados à viabilidade das aplicações que estão sendo desenvolvidas. A viabilidade dos sistemas conversacionais poderia ser comprovada por meio de estudos experimentais, como é feito por pesquisadores de engenharia de software quando buscam compreender a viabilidade de técnicas, procedimentos e novas ferramentas [23]. No entanto, trabalhos recentes apontam a falta de sistematização nos estudos que estão sendo produzidos no contexto de sistemas conversacionais [22]. Sabe-se que isso dificulta o entendimento dos procedimentos e resultados apresentados e a sua replicação [6]. Além disso, cada estudo avalia um aspecto do sistema conversacional, portanto, não há padronização no que é considerado durante a avaliação [7]. Por exemplo, enquanto alguns pesquisadores avaliam somente aspectos técnicos dos sistemas conversacionais pedagógicos, outros consideram somente as percepções dos estudantes e não observam questões técnicas, como a qualidade da conversa.

Essa falta de pesquisas sobre avaliações no contexto de sistemas conversacionais pedagógicos pode ter relação com a falta de procedimentos de apoio à sistematização, como também a falta de conhecimento sobre a necessidade dessa sistematização [8]. Como a área é constituída por pesquisadores com formações em diferentes domínios, é possível que as avaliações que estão sendo realizadas tenham relação direta com o que o pesquisador está acostumado a utilizar em trabalhos de sua área. Por exemplo, um pesquisador da área de física pode não conhecer os procedimentos que apoiam a avaliação experimental utilizados por pesquisadores da área da computação.

Neste panorama e considerando que os sistemas conversacionais se diferem dos sistemas de software tradicionais, o desafio desta pesquisa de doutorado é caracterizar os tipos de avaliação que são necessários neste contexto, propondo mecanismos para apoiar a sistematização de avaliações desses sistemas. Espera-se propor um framework de apoio para avaliações experimentais de sistemas conversacionais pedagógicos de modo que a viabilidade, qualidade e outros requisitos de interesse possam ser avaliados com maior confiabilidade. O objetivo é padronizar as avaliações experimentais de modo que os resultados possam ser melhor comparados. Para tanto, acredita-se que os procedimentos da engenharia de software experimental possam ser estendidos para o contexto dos sistemas conversacionais pedagógicos. Assim, o objetivo da pesquisa é promover direcionamento para apoiar pesquisadores a definir avaliações sistemáticas de sistemas conversacionais pedagógicos, concentrando nos diferentes estágios de uma avaliação sistemática – a saber: planejamento, operação, execução e divulgação.

2 Justificativa

Os sistemas conversacionais emergem como sistemas de software com potencial para solucionar ou minimizar problemas associados ao ensino e a aprendizagem, uma vez que flexibilizam a interação entre homem-máquina e podem estar disponíveis para uso a todo momento, isto é, 24 horas por dia e 7 dias por

semana [9]. Com base nessa premissa, muitos pesquisadores vêm tirando proveito desses sistemas para propor soluções às problemáticas que permeiam o âmbito educacional. Apenas a proposta e o desenvolvimento desses sistemas, no entanto, não garante que os problemas estão sendo solucionados.

De acordo com Shull *et al.* [18] e Travassos *et al.* [20], os produtos de software não deveriam ser apenas propostos, desenvolvidos ou apresentados para venda sem experimentação e validação. Na visão de Basili *et al.* [3], só é possível verificar se o entendimento atual sobre a temática está correto por meio de estudos experimentais. Desse modo, acredita-se que só é possível averiguar se sistemas conversacionais pedagógicos conseguem solucionar os problemas para os quais são propostos por meio da condução de avaliações/estudos experimentais.

O paradigma experimental tem sido adotado e adaptado por diversas linhas de investigação dentro da Ciência da Computação, tais como sistemas operacionais [5] e engenharia de software [23]. No âmbito da engenharia de software existem algumas definições para os propósitos da experimentação. Conradi *et al.* [4] descrevem que a experimentação auxilia a construir uma base de conhecimento confiável, permitindo reduzir as incertezas sobre quais as teorias, ferramentas e metodologias são adequadas. Adicionalmente, a experimentação elimina suposições errôneas e oferece suporte para orientar a teoria nas direções promissoras de pesquisa [4].

Apesar da experimentação estar bem estabelecida na área da engenharia de software, a condução de experimentos em outras áreas ainda não é uma prática habitual. De acordo com Kitchenham *et al.* [10], algumas áreas de pesquisa têm dificuldade em realizar estudos experimentais. Em virtude disso, existem estudos que discutem sobre a necessidade de adaptar os modelos já estabelecidos e consolidados em outras áreas ou desenvolver modelos específicos para o domínio de investigação. Considerando as necessidades intrínsecas da área e tipos de aplicações, muitas áreas têm concentrado esforços para o estabelecimento de diretrizes, processos e *frameworks* de apoio a experimentação. Na pesquisa de Sheard *et al.* [17], por exemplo, os autores mencionam sobre a necessidade de adequação de um modelo de experimentação para ambientes de apoio ao ensino de programação.

Em decorrência do atual estado da arte sobre os sistemas conversacionais pedagógicos, acredita-se que é necessário oferecer apoio metodológico, visando estimular a condução de estudos experimentais na área. Supõe-se que o apoio metodológico aumentará o rigor da pesquisa na área e contribuirá com a adesão de procedimentos sistemáticos durante a condução desses estudos. Adicionalmente, é necessário apoiar a replicação dos experimentos, estimulando a geração de pacotes de dados experimentais com conjunto de informações importantes para a replicação, buscando facilitar o compartilhamento e reúso de materiais produzidos, adotados e gerados durante a condução de experimentos.

3 Solução proposta

Considerando que o projeto de pesquisa conduzido no contexto do doutorado envolve a investigação e o estabelecimento de um *framework* para apoiar o planejamento, operação, condução e relato de estudos experimentais sobre sistemas conversacionais pedagógicos, busca-se:

(i) Estabelecer um vocabulário de experimentação para que pesquisadores conduzam e reportem seus estudos utilizando conceitos e terminologias comuns. O vocabulário é importante porque a comunidade de pesquisa que trabalha com sistemas conversacionais pedagógicos provêm de diferentes áreas do conhecimento humano e precisa conseguir utilizar a mesma linguagem para compartilhar os resultados de suas avaliações.
(ii) Definir um corpo de conhecimento sobre variáveis e métricas. O corpo de conhecimento possibilitará ao usuário final do *framework* reconhecer quais as variáveis que podem ser controladas e não podem ser controladas na condução de experimentos sobre sistemas conversacionais pedagógicos, dando ênfase às variáveis dependentes e variáveis independentes. Nessa etapa, serão investigadas e classificadas as variáveis dependentes que são de interesse para estes sistemas (*e.g.*, eficácia, desempenho, qualidade das respostas, qualidade da interface, funcionalidade, percepção e contribuição para o que se propõe). Adicionalmente, será definido, com base na literatura, um conjunto de possíveis métricas para essas variáveis.
(iii) Produzir diretrizes para que os pesquisadores ao projetarem as avaliações experimentais possam identificar as possíveis ameças à validade do estudo. Essas diretrizes devem apresentar recomendações para evitá-las.
(iv) Propor um guia de experimentação para o condução de experimentos sobre sistemas conversacionais pedagógicos. Nessa perspectiva, o guia indicará atividades que devem ser planejadas e realizadas durante a condução do experimento. O guia tem o propósito de orientar o usuário final do *framework* na realização das atividades para o planejamento e realização do experimento. Adicionalmente, o guia deverá indicar quais são as informações relevantes que devem compor o relatório do experimento.
(v) Desenvolver um ambiente de apoio a experimentação. Um ambiente que apoia a experimentação é necessário, uma vez que a condução de estudos experimentais consome tempo do investigador e produz um grande volume de informações e conhecimento [19]. Assim, buscando facilitar a condução desse tipo de experimento, uma infraestrutura informatizada deverá ser desenvolvida.

4 Procedimentos metodológicos

Para o desenvolvimento deste projeto, estão previstas oito atividades. Elas constituirão artefatos que compõem o *framework* de pesquisa:

– **Mapeamento das avaliações existentes:** inicialmente, serão localizados estudos que reportam a avaliação de sistemas conversacionais pedagógicos. Apesar desses não apresentarem avaliação sistemática, por meio deles é

possível começar a identificar variáveis e métricas que poderão fazer parte do *framework*. Para mapeá-los será replicado o mapeamento sistemático feito pelo autor e descrito em [14], incluindo critérios que permitem a seleção de estudos primários sobre sistemas conversacionais que abordem algum tipo de avaliação.

- **Definição de um vocabulário:** a replicação do mapeamento permitirá ter um entendimento mais detalhado sobre o estado real de uso de conceitos e terminologias. Assim, considerando os estudos de Wohlin *et al.* [23] e Shull *et al.* [18], será definido um vocabulário para experimentação de sistemas conversacionais pedagógicos.
- **Agrupamento de métodos e abordagens de avaliação:** um mapeamento sistemático sobre métodos e abordagens que são definidos para apoiar a avaliação de sistemas conversacionais está previsto. Espera-se que esse mapeamento possibilite a identificação de métricas, medidas e instrumentos que possam apoiar a definição do corpo de conhecimento sobre variáveis e métricas.
- **Construção do corpo de conhecimento:** Após a realização dos estudos secundários, as informações extraídas nos estudos selecionados devem ser reunidas, organizadas e disponibilizadas para acesso público como base de pesquisa e consulta. O corpo de conhecimento será definido considerando uma instanciação da abordagem de Vos *et al.* [21].
- **Reconhecimento e escrita das ameças à validade:** a definição das diretrizes para os pesquisadores reportarem ameaças à validade será inspirada em procedimentos adotados no estudo de França e Travassos [6].
- **Proposta de um guia de apoio à experimentação:** o guia para apoiar a experimentação de sistemas conversacionais pedagógicos será planejado considerando o processo experimental de Wohlin *et al.* [23]. Esse guia ilustrará procedimentos e apresentará exemplos de planejamentos de experimentos.
- **Concepção do ambiente de apoio à experimentação:** um ambiente será estabelecido buscando apoiar o gerenciamento do conhecimento científico durante todo o processo de experimentação. O ambiente será desenvolvido tomando como base o ambiente de apoio a experimentação de larga escala em engenharia de software relatado no estudo de Travassos *et al.* [19]. Adicionalmente, esse ambiente também conterá elementos que auxiliarão o pesquisador na tomada de decisão, auxiliando-o a identificar as variáveis experimentais e métricas.
- **Avaliação do framework:** após a definição dos artefatos, o *framework* de pesquisa proposto será avaliado. A avaliação será realizado por meio de um estudo experimental conduzido com pesquisadores que trabalham com sistemas conversacionais pedagógicos.

5 Resultados esperados

Com o desenvolvimento deste projeto de doutorado espera-se obter as seguintes contribuições:

1. Fortalecer o entendimento que a comunidade vinculada aos sistemas conversacionais pedagógico tem sobre avaliações sistemáticas.
2. Aprimorar a condução de estudos experimentais por intermédio de um corpo de conhecimento, um guia de experimentação e um ambiente para apoiar a automatização do processo de experimentação.
3. Possibilitar o desenvolvimento de sistemas conversacionais com mais qualidade, uma vez que a partir dos estudos experimentais é possível analisar adequadamente a viabilidade das soluções propostas.
4. Ampliar a credibilidade dos sistemas conversacionais como mecanismo de apoio ao ensino.

6 Estado atual da pesquisa

O projeto ainda está em fase inicial. Apesar disso, alguns resultados já foram obtido. Foi feito um mapeamento sistemático das pesquisas brasileiras, na tentativa de compreender como conceitos de experimentação estão sendo utilizado na avaliação dos sistemas conversacionais pedagógicos. Ao final, foram selecionados 13 estudos. Observou-se que a maioria dos estudos não define adequadamente as informações sobre os experimentos (*e.g.*, quais as variáveis investigadas). Isso gerou algumas dificuldades para reconhecer as variáveis e métricas relatadas nos estudos primários. Apesar disso, foi possível localizar uma diversidade de variáveis e métricas. O autor fez uma classificação preliminar dessas variáveis e métricas. Essa organização ainda precisa ser revisada. No próximo passo, será feita uma análise no âmbito da literatura internacional.

References

1. Abdul-Kader, S.A., Woods, J.: Survey on chatbot design techniques in speech conversation systems. International Journal of Advanced Computer Science and Applications **6**(7), 72–80 (2015)
2. Aguirre, C.C., Kloos, C.D., Alario-Hoyos, C., Muñoz-Merino, P.J.: Supporting a MOOC through a conversational agent. In: International Symposium on Computers in Education. pp. 1–6 (2018)
3. Basili, V.R., Shull, F., Lanubile, F.: Building knowledge through families of experiments. IEEE Transactions on Software Engineering **25**(4), 456–473 (1999)
4. Conradi, R., Basili, V.R., Carver, J., Shull, F., Travassos, G.H.: A pragmatic documents standard for an experience library: Roles,documen, contents and structure. Tech. rep., University of Maryland (2001)
5. Fortier, P.J., Michel, H.E.: System performance evaluation tool selection and use. In: Fortier, P.J., Michel, H.E. (eds.) Computer Systems Performance Evaluation and Prediction, pp. 331 – 344. Digital Press, Burlington (2003)
6. França, B.B.N., Travassos, G.H.: Experimentation with dynamic simulation models in software engineering: planning and reporting guidelines. Empirical Software Engineering **21**(3), 1302–1345 (2016)
7. Hobert, S.: How are you, chatbot? evaluating chatbots in educational settings–results of a literature review. In: Lecture Notes in Informatics, pp. 259–270 (2019)

8. Io, H., Lee, C.: Chatbots and conversational agents: A bibliometric analysis. In: International Conference on Industrial Engineering and Engineering Management. pp. 215–219 (2017)
9. Jain, M., Kumar, P., Kota, R., Patel, S.N.: Evaluating and informing the design of chatbots. In: Designing Interactive Systems Conference. pp. 895–906 (2018)
10. Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., Emam, K.E., Rosenberg, J.: Preliminary guidelines for empirical research in software engineering. IEEE Transactions on Software Engineering **28**(8), 721–734 (2002)
11. Krassmann, A.L., Paz, F.J., Silveira, C., Tarouco, L.M.R., Bercht, M.: Conversational agents in distance education: comparing mood states with students' perception. Creative Education **9**(11), 1726–1742 (2018)
12. Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y., et al.: Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association **25**(9), 1248–1258 (2018)
13. Montenegro, J.L.Z., da Costa, C.A., da Rosa Righi, R.: Survey of conversational agents in health. Expert Systems with Applications **129**, 56 – 67 (2019)
14. Paschoal, L.N.: Contribuições ao ensino de teste de software com o modelo flipped classroom e um agente conversacional. Master's thesis, USP (2019)
15. Paschoal, L.N., de Oliveira, M.M., Chicon, P.M.M.: A chatterbot sensitive to student's context to help on software engineering education. In: Latin American Computer Conference (CLEI). pp. 839–848 (2018)
16. Paschoal, L.N., Turci, L.F., Conte, T.U., Souza, S.R.S.: Towards a conversational agent to support the software testing education. In: Brazilian Symposium on Software Engineering. pp. 57–66 (2019)
17. Sheard, J., Simon, S., Hamilton, M., Lönnberg, J.: Analysis of research into the teaching and learning of programming. In: International workshop on Computing education research. pp. 93–104 (2009)
18. Shull, F., Carver, J., Travassos, G.H.: An empirical methodology for introducing software processes. In: International symposium on Foundations of software engineering. pp. 288–296 (2001)
19. Travassos, G.H., d. Santos, P.S.M., Mian, P.G., Neto, A.C.D., Biolchini, J.: An environment to support large scale experimentation in software engineering. In: International Conference on Engineering of Complex Computer Systems. pp. 193–202 (2008)
20. Travassos, G.H., Gurov, D., Amaral, E.A.G.: Introdução à engenharia de software experimental. Tech. rep., Rio de Janeiro (2002)
21. Vos, T.E.J., Marín, B., Escalona, M.J., Marchetto, A.: A methodological framework for evaluating software testing techniques and tools. In: ICQS. pp. 230–239 (2012)
22. Winkler, R., Söllner, M.: Unleashing the potential of chatbots in education: A state-of-the-art analysis. In: Academy of Management Annual Meeting. pp. 1–40 (2018)
23. Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in software engineering. Springer Science & Business Media (2012)
24. Xu, A., Liu, Z., Guo, Y., Sinha, V., Akkiraju, R.: A new chatbot for customer service on social media. In: Conference on Human Factors in Computing Systems. pp. 3506–3510 (2017)