Un système d'aide au dialogue en santé intime des femmes

Xingyu Liu, François Portet, Didier Schwab, Juliette Mauro

Un système d'aide au dialogue en santé intime des femmes

Xingyu Liu[1,2], François Portet[1], Didier Schwab[1], Juliette Mauro[2]

[1] Univ. Grenoble Alpes, CNRS, Grenoble INP*, LIG, 38000 Grenoble, France
[2] Shesmet

xingyu.liu@univ-grenoble-alpes.fr

Résumé

Dans le cadre de notre projet visant à fournir des conseils professionnels efficaces sur la santé intime et sexuelle des femmes sur une plateforme numérique, nous avons pour objectif de créer un système d'assistance de dialogue pour aider les experts gynécologues à répondre aux questions des utilisatrices. Ce système est spécifiquement conçu pour proposer une liste de réponses possibles que l'expert peut sélectionner, permettant ainsi d'optimiser le temps et de se concentrer sur la qualité des soins apportés. Nous présentons dans cet article la modélisation de ce système ainsi qu'un module de recherche basé sur la recherche d'informations, qui utilise la similarité sémantique pour suggérer des réponses pertinentes.

Mots-clés

Santé intime des femmes, Système d'aide au dialogue, Recherche d'information

Abstract

In order to efficiently provide professional advice on women's intimate and sexual health in a digital platform, we are undertaking a project aimed at creating a dialogue assistance system to assist gynecological experts in responding to user questions. The system is designed to provide a list of possible responses for the expert to choose from, saving time and allowing them to focus on care. The article presents the system's modeling and a search module based on information retrieval that uses semantic similarity to suggest relevant responses.

Keywords

Women's intimate health, Dialogue assistance system, Information retrieval

1 Introduction

La santé intime des femmes est un sujet encore peu abordé dans son ensemble et souvent résumé à la santé reproductive [1, 8]. Pourtant les femmes ont physiologiquement plusieurs étapes de vie qui vont impacter de manière plus ou moins forte leur bien-être mental et physique : la puberté, la ménopause et l'après-ménopause. Certaines vont également connaître la maternité. L'accès à une information de qualité, personnalisée et anonymisée représente un fort vecteur d'autonomisation et d'égalité de soins pour l'ensemble de la population féminine. Pourtant, aujourd'hui les femmes voulant se renseigner sur ces thèmes sont souvent en prise avec un flot d'informations qui peuvent être discordantes, incomplètes et de sources non vérifiables, par exemple, dans les forums de santé alimentés par les utilisateurs.

Dans ce contexte, la société Shesmet, spécialiste dans le domaine de la santé intime des femmes, a lancé une application MySLife [1] visant à connecter les femmes autour de la santé intime et sexuelle et à établir des échanges entre les utilisatrices de l'application et les experts gynécologiques à travers des dialogues textuels. Dans le cadre d'une collaboration CIFRE avec Shesmet, notre projet a pour objectif de concevoir un système d'aide au dialogue textuel pour les experts en santé intime de la femme. Plus concrètement, dans l'application MySLife, quand l'utilisatrice lance un échange privé avec l'expert·e, après chaque tour de parole de l'utilisatrice, le système doit proposer une liste des réponses possibles à suggérer à l'expert·e que celui-ci sélectionnera. Un tel système permet d'associer les informations pertinentes à la question de l'utilisatrice. L'expert·e peut ainsi directement choisir les réponses fournies par le système ou éventuellement les modifier. Cela économise le temps de rédaction des réponses de l'expert·e, et l'expert·e peut ainsi se focaliser sur les soins.

Dans cet article, nous présentons dans la section 2 la modélisation du système de dialogue en détaillant le fonctionnement du système d'aide au dialogue dans un tel environnement. Nous présentons ensuite le travail actuellement développé sur le module de recherche de réponse dans la section 3. Nous tenons à donner des arguments en faveur d'un système d'aide au dialogue autour de l'aide, les défis pour créer un tel système, et les perspectives futures de notre projet en considérant l'arrivée de ChatGPT.

2 Modélisation du système de dialogue

Le système que nous proposons n'a pas d'équivalent dans le domaine de la santé à notre connaissance. Ceci pose la question de comment modéliser les échanges entre les différents acteurs lors du dialogue.

*Institute of Engineering Univ. Grenoble Alpes

1. https ://myslife.co/mobile/

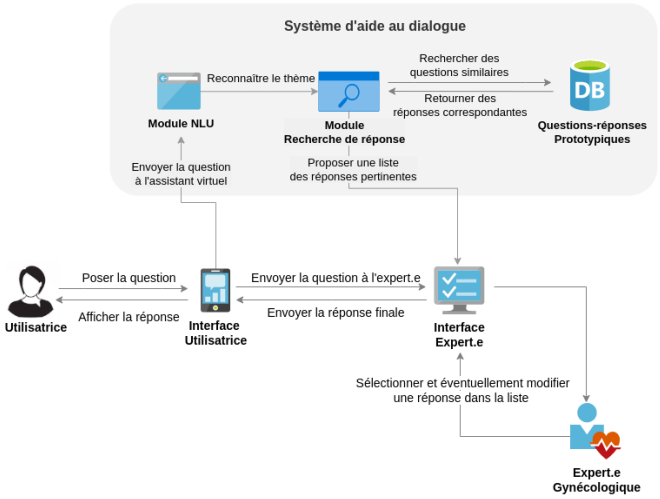

FIGURE 1 – Schéma des étapes globales du système de dialogue

La figure 1 schématise le système de dialogue comme envisagé dans son écosystème. Il se compose de trois agents : l'utilisatrice, le système d'aide au dialogue et l'expert·e gynécologique.
Les éléments que nous avons dans l'environnement sont : l'interface utilisatrice (l'application MySLife sur le téléphone portable), le module NLU qui extrait les informations de la requête de l'utilisatrice, le modèle de recherche des questions similaires, la base de données des questions-réponses prototypiques, et l'interface experte pour que l'expert·e puisse choisir ou modifier une des réponses prototypiques proposées par le système d'aide au dialogue.
Le processus complet d'un tour de question-réponse est comme suit :

1. l'utilisatrice pose sa question dans l'application ;
2. cette question est envoyée au système d'aide au dialogue et à l'interface de l'expert·e ;
3. le module NLU reconnaît la catégorie de la question (endométriose, douleur au rapport, ménopause, etc.) ;
4. des questions similaires sont recherchées dans la base de données prototypiques et une liste de réponses prototypiques est fournie à l'expert·e à travers l'interface dédiée ;
5. l'expert·e choisit une ou des réponses, les modifie si nécessaire puis envoie la réponse finale à l'interface de l'utilisatrice.

3 Module de recherche de réponse basé sur la recherche d'information

Pour concevoir un assistant qui propose une liste de réponses à l'expert·e, nous menons d'abord des expériences concernant le module de recherche de réponse.

Nous explorons la performance d'une approche basée sur la recherche d'information qui utilise la similarité sémantique entre la question et les questions prototypiques, stockées dans la base de données prototypiques, afin d'apparier à une question donnée, sa réponse prototypique correspondante.
En effet, la représentation sémantique des questions est variable pour les systèmes de question-réponse basés sur la recherche d'information. L'une des premières approches est d'utiliser le modèle de sac de mots qui ordonne les documents en fonction de la fréquence des termes qui apparaissent dans chaque document, comme BM25 [10]. Avec les progrès de l'apprentissage profond, les modèles word2vec entraînés sur un corpus spécifique pour la représentation des questions [7] ou encore les modèles RNN pour encoder les conversations en embeddings [2] ont été de plus en plus utilisés. Récemment, un large éventail d'études ont montré que des modèles de langage pré-entraînés sur un large corpus peuvent apprendre des représentations de langage universelles. Certaines études dans le domaine du système de dialogue proposent d'apprendre le modèle d'appariement contexte-réponse avec des tâches auxiliaires auto-supervisées conçues pour les données de dialogue basées sur des modèles de langage pré-entraînés (par exemple, BERT) [11].
La figure 2 montre le pipeline de l'approche de sélection de réponses. Le calcul de similarité se réalise par le cosinus des plongements de la question de l'utilisatrice (Q_u) et des questions prototypiques (Q_{p1}, Q_{p2} ...). Nous prenons les trois questions prototypiques les plus similaires à la Q_u, et retournons les trois réponses correspondantes.

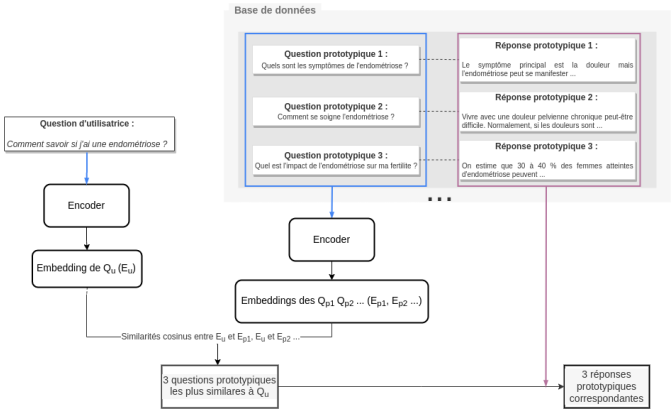

FIGURE 2 – Pipeline de l'approche de sélection de réponse

3.1 Méthodes du plongement de phrases

Afin d'avoir les représentations des questions par vecteurs dans un espace multidimensionnel, nous avons testé trois méthodes du plongement de phrases, qui sont respectivement la moyenne pondérée des vecteurs des mots de la question, le plongement de la question par le modèle de Sentence-BERT (SBERT) [9] et le plongement de la question par le modèle multilingue Universal Sentence Encoder (USE) [12].
Concrètement, pour apprendre une représentation vectorielle des mots, nous avons utilisé le module FastText [3].

Nous avons entraîné les plongements avec un corpus Doctissimo du forum français Doctissimo sur la santé, et plus particulièrement de sa catégorie "Gynécologie - Santé de la femme" (491738 paires de questions-réponses).
Les paramètres que nous avons choisi pour l'entraînement sont : 300 pour la taille des vecteurs, 5 pour la taille de la fenêtre de contexte, et le modèle ignore tous les mots dont la fréquence totale est inférieure à 5.
Par la suite, la manière que nous avons choisi pour obtenir une représentation vectorielle des questions est d'effectuer une certaine arithmétique vectorielle sur tous les vecteurs correspondant aux mots du document pour les résumer en un seul vecteur dans le même espace de plongement de mots. Nous avons utilisé l'opérateur de la moyenne. De manière formelle, celle-ci peut être représentée comme dans l'équation 1 :

$$V = \frac{1}{i} \sum_{k=1}^{i} v_k \tag{1}$$

où v_k est le vecteur du k^e mot du document.
Ultérieurement, afin d'améliorer l'identification des mots qui sont plus importants dans chaque phrase, nous avons adopté la méthode de pondération proposée dans [4]. Des pondérations dépendant de la fréquence inverse en document ont été appliquées sur les phrases examinées. Ainsi, la représentation vectorielle d'une phrase devient alors :

$$V = \frac{1}{i} \sum_{k=1}^{i} v_k * idf(w_k) \tag{2}$$

où v_k est le vecteur du k^e mot du document et $idf(w)$ la fonction qui donne l'IDF du mot w.
Les modèles SBERT et USE font tous les deux correspondre leur entrée à des vecteurs dans un espace vectoriel de 512 dimensions. Nous avons utilisé les modèles pré-entraînés multilingues [2] de SBERT et d'USE [3], couvrant respectivement 15 langues et 16 langues, dont le français.
Le modèle multilingue d'USE construit des plongements de phrases en utilisant le sous-graphe d'encodage de l'architecture du transformer. L'encodeur utilise le mécanisme d'attention pour calculer des représentations contextuelles des mots d'une phrase qui prennent en compte à la fois l'ordre des mots et le contexte environnant. Les représentations contextuelles des mots sont moyennées ensemble pour obtenir un plongement au niveau de la phrase.
Les implémentations de SBERT et d'USE sont différentes. SBERT est basé sur PyTorch, et USE sur TensorFlow.

3.2 Jeux de données d'évaluation

Nous avons effectué deux évaluations, une pour évaluer les trois méthodes du plongement de phrases indépendamment de l'application, et l'autre pour évaluer la performance de notre modèle sur un jeu de données liées à l'application visée par le projet.

Pour la première évaluation, le corpus CLISTER, un corpus pour la similarité sémantique textuelle dans des cas cliniques généraux en français, proposé par [6] a été utilisé pour évaluer la performance des méthodes d'embeddings sur la similarité sémantique. Le corpus CLISTER contient 1 000 paires de phrases sélectionnées aléatoirement depuis le corpus CAS [5] (un corpus qui contient des cas cliniques publiés dans la littérature scientifique et du matériel de formation en français avec des annotations sémantiques.), et annotées manuellement en scores de similarité. Nous avons entraîné le modèle FastText sur le jeu de données d'entraînement du corpus CLISTER, soit 600 paires de phrases, et pour les modèles USE et SBERT, nous les avons utilisé sans ajustement. L'évaluation des trois méthodes d'embeddings s'est effectuée sur le jeu de données de test du corpus CLISTER, soit 400 paires de phrases.
Quant à la deuxième évaluation, nous utiliserons notre propre corpus d'évaluation. Ce corpus est initialement composé de 18 paires de questions-réponses écrites par l'experte gynécologique de l'entreprise Shesmet et de 10 paires de questions-réponses proprement rédigées sélectionnées depuis le corpus MedDialog [13]. Ensuite, pour augmenter le nombre d'exemples dans le jeu d'évaluation, basé sur ces 28 paires de prototypes en français, nous avons adopté les méthodes de traduction aller-retour (round-trip translation) et de remplacement des mots par synonymes afin d'augmenter les données d'évaluation. Nous avons utilisé l'outil de traduction automatique DeepL [4] pour les traductions aller-retour d'anglais-chinois et d'anglais-turc. Pour la méthode de remplacement des mots par synonyme, le modèle FastText avec ajustement sur le corpus Doctissimo a été utilisé. L'ensemble de notre jeu de données d'évaluation contient finalement 100 paires de questions-réponses prototypiques.

3.3 Métriques d'évaluation

Pour l'évaluation des trois méthodes de plongement de phrases, nous avons adopté la méthode présentée dans l'étude de [6]. Nous avons évalué la similarité sur les données de test de CLISTER à l'aide de la corrélation de Spearman.
Ensuite, pour évaluer la performance des trois méthodes du plongement de phrases, nous avons choisi deux métriques : accuracy et Mean Reciprocal Rank (MRR). Pour décider si une réponse retournée était correcte, la réponse devait respecter deux contraintes :

1. la réponse correspondante dans le corpus d'évaluation se trouve dans la liste de réponses retournées par le système ;
2. le score de similarité entre la réponse retournée et la réponse d'évaluation dépasse 0.5.

L'*accuracy* est ensuite calculée comme étant le nombre de bonnes réponses sur le nombre total d'exemples dans le jeu de données d'évaluation.
Le score MRR est une mesure permettant d'évaluer les systèmes qui renvoient une liste classée de réponses à des re-

2. distiluse-base-multilingual-cased-v1
3. universal-sentence-encoder-multilingual-large
4. https ://www.deepl.com/

Méthode	Données de test	Spearman
FastText entraîné sur Clister-train	Clister-test	0.57
Universal Sentence Encoder	Clister-test	0.78
Sentence-Bert	Clister-test	0.75

TABLE 1 – Résultats des expériences de similarité sémantique avec différents méthodes d'embeddings

Méthode	Données de test	Accuracy	MRR
FastText entraîné sur Doctissimo	eval-100	0.56	0.82
Universal Sentence Encoder	eval-100	0.37	0.83
Sentence-Bert	eval-100	0.50	0.79

TABLE 2 – Résultats des performances du modèle de sélection de réponses avec différents méthodes d'embeddings

quêtes. Pour une seule réponse, le rang réciproque est la position de la réponse correcte la mieux classée renvoyée dans une liste de réponses. Pour Q réponses multiples, comme dans notre cas, le MRR est la moyenne des rangs réciproques (cf. équation (3)).

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rang_i} \tag{3}$$

3.4 Résultats

Le tableau 1 présente les scores de Spearman de différentes méthodes du plongement de phrases pour l'évaluation sur le corpus CLISTER. Le meilleur score Spearman pour cette tâche a été obtenu par la méthode USE avec un score de 0,78. Ces résultats suggèrent que USE est plus performant pour capturer la similarité sémantique entre les paires de phrases dans le corpus CLISTER par rapport à Sentence-BERT et FastText. Les méthodes USE et SBERT montrent leur capacité à capturer des informations sémantiques plus riches et contextualisées par rapport à FastText qui utilise des modèles de sacs de mots et des représentations de mots basées sur la fréquence.

Les résultats de l'évaluation sur notre corpus de test sont résumés dans le tableau 2. Les scores MRR sont relativement bons pour toutes les trois méthodes, ce qui indique que si la bonne réponse a été trouvée, elle est souvent au premier ou deuxième rang, c'est-à-dire qu'elle a un score de similarité relativement haut.

Les scores d'accuracy sont faibles pour les trois méthodes. Étant donné que nous n'avions pas de données pour ajuster les modèles USE et SBERT, il est possible qu'ils ne soient pas adaptés à nos données, celles-ci contenant une part importante de termes médicaux dont entre autres les termes sur la santé intime des femmes. Une piste d'amélioration est de créer un jeu de données spécifique à la santé intime des femmes avec lequel nous pouvons ajuster USE et SBERT. Pour la méthode de FastText, le corpus Doctissimo lui a fourni un espace sémantique de la santé intime des femmes relativement grand avec 118,956 tokens uniques, ce qui pourrait entraîner de meilleures performances dans ce domaine particulier par rapport à USE et SentenceBERT. D'ailleurs, nous avons observé que les erreurs se trouvaient souvent dans les exemples longs. Des améliorations ultérieures peuvent être apportées en essayant de trouver une meilleure représentation pour les phrases longues.

4 Discussion et travaux futurs

L'objectif final de ce travail de recherche est de créer un système d'aide au dialogue qui propose des réponses pré-établies en fonction de la question de l'utilisatrice et du contexte dialogique pour les experts gynécologiques. L'intérêt d'un tel système est d'abord de diminuer le temps de rédaction des réponses de l'expert·e. De plus, le système permet à l'utilisatrice de l'application ayant des questionnements sur la santé intime de bénéficier d'une interaction plus efficace.

La mise au point de ce système soulève de nombreuses questions que nous explorerons dans le futur.

En premier lieu, les approches par apprentissage automatique nécessitent de grands corpus équilibrés pour garantir une performance optimale, alors que l'accès aux données dans le domaine de la santé reste toujours un grand défi, notamment pour le domaine de la santé intime des femmes, un nouveau domaine qui est dénué de corpus. Nous travaillons en ce moment sur notre corpus sur la santé intime des femmes en français. Ce corpus sera disponible d'ici fin juin.

Dans un deuxième temps, l'entreprise possède une expertise dans le domaine cible, il sera donc nécessaire de s'interroger sur la manière dont la connaissance métier pourra être utilisée pour guider le dialogue.

La sortie récente de ChatGPT a suscité un grand intérêt auprès du public et des professionnels de la santé. Nous étudions actuellement ses points forts, ses limites, en particulier éthiques, en ce qui concerne les soins de santé en particulier pour la santé intime des femmes. Les analyses et résultats de nos travaux feront l'objet d'une publication future.

Remerciements

Ce travail est effectué dans le cadre d'une convention CIFRE, gérée par l'Association Nationale de la Recherche Technique (ANRT), et établie entre le Laboratoire d'informatique de Grenoble et la société Shesmet.

Références

[1] N. Bajos, F. Prioux, and C. Moreau. L'augmentation du recours répété à l'ivg en france : des enjeux contraceptifs au report de l'âge à la maternité. *Revue d'épidémiologie et de santé publique*, 61(4) :291–298, 2013.

[2] A. Bartl and G. Spanakis. A retrieval-based dialogue system utilizing utterance and context embeddings. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1120–1125. IEEE, 2017.

[3] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information.

Transactions of the association for computational linguistics, 5 :135–146, 2017.

[4] J. Ferrero, D. Schwab, et al. Amélioration de la similarité sémantique vectorielle par méthodes non-supervisées. In *24e conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017)*, 2017.

[5] N. Grabar, V. Claveau, and C. Dalloux. Cas : French corpus with clinical cases. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 122–128, 2018.

[6] N. Hiebel, O. Ferret, K. Fort, and A. Névéol. Clister : A corpus for semantic textual similarity in french clinical narratives. In *LREC 2022-International Conference on Language Resources and Evaluation (LREC)*, 2022.

[7] N. Othman, R. Faiz, and K. Smaïli. Using word embeddings to retrieve semantically similar questions in community question answering. *Journal of International Science and General Applications*, 1(1), 2018.

[8] L. Poncet. *Santé sexuelle et reproductive des femmes migrantes sans logement hébergées à l'hôtel en Ile-de-France*. PhD thesis, université Paris-Saclay, 2021.

[9] N. Reimers and I. Gurevych. Sentence-bert : Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv :1908.10084*, 2019.

[10] W. Sakata, T. Shibata, R. Tanaka, and S. Kurohashi. Faq retrieval using query-question similarity and bert-based query-answer relevance. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1113–1116, 2019.

[11] R. Xu, C. Tao, D. Jiang, X. Zhao, D. Zhao, and R. Yan. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14158–14166, 2021.

[12] Y. Yang, D. Cer, A. Ahmad, M. Guo, J. Law, N. Constant, G. H. Abrego, S. Yuan, C. Tar, Y.-H. Sung, et al. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv :1907.04307*, 2019.

[13] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, et al. Meddialog : Large-scale medical dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.