Concevoir un assistant conversationnel de manière itérative et semi-supervisée avec le clustering interactif

Erwan Schild, Gautier Durantin, Jean-Charles Lamirel

Concevoir un assistant conversationnel de manière itérative et semi-supervisée avec le clustering interactif

Erwan Schild[*,**], Gautier Durantin[*], Jean-Charles Lamirel[**]

[*] Euro-Information Développements, Groupe Crédit-Mutuel
4 Rue Frédéric-Guillaume Raiffeisen 67000 Strasbourg,
prenoms.nom@e-i.com, https://www.e-i.com/
[**] LORIA, Université de Lorraine
615 Rue du Jardin-Botanique, 54506 Vandoeuvre-lès-Nancy,
prenoms.nom@loria.fr, https://www.loria.fr/

Résumé. La création d'un jeu de données nécessaire à la conception d'un assistant conversationnel résulte le plus souvent d'une étape manuelle et fastidieuse qui manque de techniques destinées à l'assister. Pour accélérer cette étape d'annotation, nous proposons une méthode de *clustering* interactif : il s'agit d'une approche itérative inspirée de l'apprentissage actif, reposant sur un algorithme de *clustering* et tirant parti d'une annotation de contraintes pour guider le regroupement des questions en une structure d'intentions. Dans cet article, nous exposons la méthodologie à mettre en oeuvre pour concevoir un assistant conversationnel opérationnel à l'aide du *clustering* interactif.

1 Introduction et enjeux

L'utilisation des assistants conversationnels (*chatbot*) explose car ces derniers permettent efficacement d'accéder à l'information avec des requêtes en langage naturel. Toutefois, leur création est encore fastidieuse en raison du manque de techniques permettant d'assister l'annotation du jeu de données nécessaire à leur entrainement. En effet, cette étape résulte le plus souvent d'un travail manuel et empirique possédant plusieurs défauts. On peut notamment citer le besoin de définir un modèle de catégorisation des données en intentions [1] ou encore la difficulté de maintenir cette modélisation abstraite sans introduire de biais ou d'ambiguïtés.

Pour assister l'humain dans cette tâche d'annotation, nous cherchons une alternative à l'approche manuelle en introduisant des initiatives de la machine. Nous nous intéressons plus particulièrement à la méthode du *clustering* interactif définie initialement par Gançarski et Wemmert (2007) et Lampert et al. (2019) pour la détection d'objets dans une image, une approche reposant sur la combinaison entre l'apprentissage actif et le *clustering* sous contraintes.

Néanmoins, le *clustering* interactif reste une méthode peu explorée dans le cadre du traitement du langage naturel. Dans Schild et al. (2021), nous avons proposé une première implémentation fonctionnelle cette technique, et nous nous intéressons désormais à l'intégration du *clustering* interactif dans le processus de conception d'un assistant conversationnel. Pour cet

1. Par exemple, «*Joue-moi du jazz !*» peut être modélisé par l'intention "*jouer de la musique*" et l'entité "*jazz*".

article, nous ferons abstraction des estimations de charge de travail pour discuter des impacts méthodologiques de cette nouvelle approche et définir les flux de validation associés. Ces estimations de charge seront réalisées dans une étude ultérieure mettant en oeuvre le protocole défini.

2 Clustering interactif

Principe général Le *clustering* interactif est une méthode semi-supervisée qui repose sur l'alternance successive entre une annotation de contraintes (*MUST-LINK*, *CANNOT-LINK*) et un *clustering* sous contraintes. L'objectif ainsi recherché est la création d'un cercle vertueux (cf. figure 1) pour améliorer itérativement la pertinence du *clustering* obtenu : un oracle suggère un ensemble de contraintes à annoter pour corriger efficacement le *clustering* issu de l'itération précédente, et le *clustering* exploite les contraintes annotées jusqu'à présent pour suggérer un nouveau partitionnement plus pertinent pour l'itération suivante.

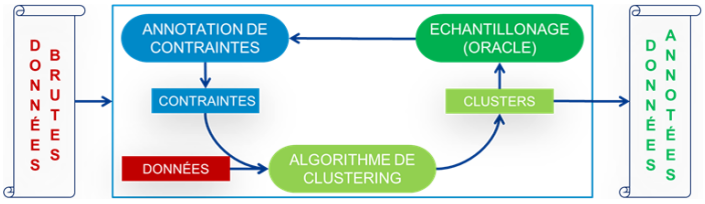

FIG. 1 – *Schéma représentant la boucle d'itérations du clustering interactif. Les étapes en bleu représentent l'initiative humaine, et les étapes en vert représentent l'initiative machine.*

Résultats obtenus Dans Schild et al. (2021), nous avons, à l'aide d'un corpus de données textuelles issues du domaine bancaire, confirmé la faisabilité technique d'une annotation avec cette méthode, nous permettant ainsi de déduire les propriétés suivantes :

- la définition d'un modèle d'organisation des données en intentions n'est plus un prérequis pour réaliser la phase d'annotation : en effet, cette structure d'intention émerge naturellement des contraintes annotées au cours des itérations, offrant ainsi un gain de temps majeur pour la conception d'un assistant ;
- avec l'utilisation de contraintes, l'annotation devient un mécanisme binaire centré sur la similarité des réponses à donner aux questions : l'annotateur n'a donc plus besoin d'avoir une connaissance globale de la structure d'intentions définie, ce qui réduit la complexité de sa tâche et minimise la possibilité d'introduire des contradictions ;
- la tâche d'annotation est désormais partagée entre l'homme et la machine : la charge de travail de l'annotateur est donc réduite car il n'intervient que pour fournir la dose d'information nécessaire pour améliorer la pertinence du *clustering* obtenu ;
- au cours des itérations, le partitionnement des données peut être efficacement corrigé grâce à l'annotation des contraintes adéquates : le choix de la méthode de sélection des données à annoter est donc très important ;
- comme les interactions successives permettent d'améliorer le partitionnement des données, il est possible d'obtenir un résultat pertinent malgré l'emploi d'algorithmes de

clustering simples : on peut donc privilégier la rapidité à la performance pour choisir la méthode de *clustering* car le mécanisme d'itérations comblera en partie cette lacune.

3 Intégration pour la conception d'un chatbot complet

Pour utiliser le *clustering* interactif dans le cycle de conception d'un assistant conversationnel, nous devons définir un mode d'emploi pour s'assurer de la pertinence du partitionnement des données que l'on obtient. Le protocole d'utilisation que nous proposons est schématisé dans la figure 2 et est détaillé ci-après. Ce dernier met en avant plusieurs pistes d'étude que nous devrons traiter ultérieurement afin de confirmer la viabilité de ce protocole.

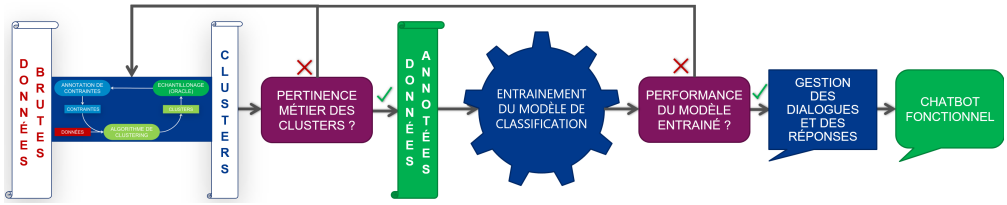

FIG. 2 – *Schéma représentant le protocole de mise en oeuvre d'un assistant conversationnel avec le clustering interactif. Les étapes en bleu foncé sont les actions à réaliser, celles en violet sont les évaluations intermédiaires et les objets en vert sont les livrables de la méthode.*

Annotation avec le *clustering* interactif. Bien que l'implémentation de notre méthode soit fonctionnelle pour le traitement de données textuelles, il reste quelques inconnues à résoudre pour la rendre pleinement opérationnelle.

Tout d'abord, on peut se demander comment estimer le nombre de *clusters* optimal. En effet, cette information est initialement inconnue et difficilement identifiable. Un *clustering* collaboratif pourrait représenter une piste pour se rapprocher de ce nombre optimal au cours des itérations, voire pour l'obtenir.

Ensuite, il faut être capable de prévenir ou résoudre un conflit d'annotation de contraintes entre les données. Ce problème est bien connu de l'apprentissage incrémental, et une étude approfondie de ce type d'apprentissage devrait permettre de le contourner.

Enfin, il demeure la question de l'exhaustivité de l'annotation. En effet, annoter toutes les contraintes entre les données est impensable en pratique compte tenu de la charge de travail nécessaire. Il peut être néanmoins pertinent de se contenter d'une annotation partielle et d'estimer si le jeu de données résultant est suffisamment fiable pour concevoir un assistant. Les paragraphes suivants détaillent les analyses qu'il est possible de mener en ce sens.

Analyse de la pertinence sémantique. Après plusieurs itérations du *clustering* interactif, nous pouvons nous interroger sur la pertinence sémantique des *clusters* obtenus : pouvons-nous associer une intention à ces *clusters* ? Comme l'annotation de contraintes se fait sur la base de similarité de la réponse (Schild et al., 2021), il est légitime de supposer que chaque

cluster doit en effet pouvoir être identifié par une réponse générale. Ce critère est fortement discriminant car il conditionne la connaissance dont l'assistant dispose. En conséquence, un *clustering* jugé trop peu pertinent demandera l'annotation de contraintes supplémentaires.

Analyse de la performance statistique. Si l'analyse de la pertinence des *clusters* est satisfaisante, nous pouvons alors nous intéresser aux contraintes techniques liées à la classification des intentions. En effet, la qualité de l'assistant dépend en grande partie de l'efficacité de cette détection, et il faut donc s'assurer de la performance du modèle entraîné, par exemple avec l'emploi d'un test K-folds et l'analyse d'une matrice de confusion. Si les performances ne sont pas satisfaisantes, l'analyse des dépendances entre *clusters* permettrait de définir si un remaniement manuel de la structure d'intention suffirait à corriger cette situation ou s'il faudrait ré-annoter d'autres contraintes pour identifier des intentions non reconnues jusque-là.

4 Conclusion

En suivant les perspectives que nous avons décrites, et après validation des critères de pertinence et de performance, il ne resterait qu'à définir les réponses à assigner à chaque intention pour compléter l'assistant conversationnel. Avec la définition de notre phase d'affectation des contraintes, cette dernière étape serait par ailleurs triviale. Nous arriverions ainsi au terme de la procédure, et nous disposerions donc au final d'un assistant conversationnel fonctionnel grâce à une approche semi-supervisée dont les bases ont déjà été définies dans Schild et al. (2021).

Références

Gançarski, P. et C. Wemmert (2007). Collaborative multi-step mono-level multi-strategy classification. *Multimedia Tools and Applications 35*(1), 1–27.

Lampert, T., B. Lafabregue, et P. Gançarski (2019). Constrained Distance based K-Means Clustering for Satellite Image Time-Series. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2419–2422. IEEE.

Schild, E., G. Durantin, J.-C. Lamirel, et F. Miconi (2021). Conception itérative et semi-supervisée d'assistants conversationnels par regroupement interactif des questions. *EGC 2021*.

Summary

The design of a dataset needed to train a chatbot is most often the result of manual and tedious step. To guarantee the efficiency of the annotation, we propose the interactive clustering method, an active learning method based on constraints annotation. It's an iterative approach, relying on a constrained clustering algorithm and using annotator knowledge to lead clustering. In this paper, we expose the process to design a chatbot with the interactive clustering method.