# 1. ABSTRACT CLASSIQUE (en moins de 4000 caractères)

## 1.A. En français

La tâche d'annotation, nécessaire à l'entraînement d'assistants conversationnels, fait habituellement appel à des experts du domaine à modéliser. Toutefois, l'annotation de données est connue pour être une tâche difficile en raison de sa complexité et sa subjectivité : elle nécessite par conséquent de solides compétences analytiques dans le but de modéliser les textes en intention de dialogue. De ce fait, la plupart des projets d'annotation choisissent de former les experts aux tâches d'analyse pour en faire des "super-experts".

Dans cette thèse, nous avons plutôt décidé mettre l’accent sur les connaissances réelles des experts en proposant une nouvelle méthode d'annotation basée sur un Clustering Interactif. Celle-ci se base sur une coopération Homme/Machine, où la machine réalise un clustering pour proposer une base initiale d'apprentissage, et où l'expert annote des contraintes MUST-LINK ou CANNOT-LINK entre les données pour affiner itérativement la base d'apprentissage proposée. Une telle annotation présente l'avantage d'être plus instinctive, car les experts peuvent associer ou différencier les données en fonction de la similarité de leur cas d'usage, permettant ainsi de traiter les données comme ils le feraient professionnellement au quotidien.

Au cours de nos études, nous avons pu montrer que cette méthode diminuait sensiblement la complexité de conception d'une base d'apprentissage, réduisant notamment la nécessité de formation des experts intervenant dans un projet d'annotation. Nous proposons une implémentation technique de cette méthode (algorithmes et interface graphique associée), ainsi qu'une étude des paramètres optimaux pour obtenir une base d'apprentissage cohérente en un minimum d'annotation. Nous réalisons également une étude de coûts (techniques et humains) permettant de confirmer que l'utilisation d'une telle méthode est réaliste dans un cadre industriel. De plus, afin que la méthode atteigne son plein potentiel, nous fournissons un ensemble de conseils, notamment : (1) des recommandations visant à cadrer la stratégie d’annotation, (2) une aide à l'identification et à la résolution des divergences d'opinion entre annotateurs, (3) des indicateurs de rentabilité pour chaque intervention de l'expert, et (4) des méthodes d'analyse de la pertinence de la base d'apprentissage en cours de construction.

En conclusion, cette thèse offre une approche innovante pour concevoir une base d'apprentissage d'un assistant conversationnel, permettant d'impliquer les experts du domaine métier pour leurs vraies connaissances, tout en leur demandant un minimum de compétences analytiques et techniques. Ces travaux ouvrent ainsi la voie à des méthodes plus accessibles pour la construction de ces assistants.

## 1.B. In English

Usually, the task of annotation, used to train conversational assistants, relies on domain experts who understand the subject matter to model. However, data annotation is known to be a challenging task due to its complexity and subjectivity. Therefore, it requires strong analytical skills to model the text in dialogue intention. As a result, most annotation projects choose to train experts in analytical tasks to turn them into "super-experts".

In this thesis, we decided instead to focus on the real knowledge of experts by proposing a new annotation method based on Interactive Clustering. This method involves a Human-Machine cooperation, where the machine performs clustering to provide an initial learning base, and the expert annotates MUST-LINK or CANNOT-LINK constraints between the data to iteratively refine the proposed learning base. Such annotation has the advantage of being more instinctive, as experts can associate or differentiate data according to the similarity of their use cases, allowing them to handle the data as they would professionally do on a daily basis.

During our studies, we have been able to show that this method significantly reduces the complexity of designing a learning base, notably by reducing the need for training the experts involved in an annotation project. We provide a technical implementation of this method (algorithms and associated graphical interface), as well as a study of optimal parameters to achieve a coherent learning base with minimal annotation. We have also conducted a cost study (both technical and human) to confirm that the use of such a method is realistic in an industrial context. Finally, we provide a set of recommendations to help this method reach its full potential, including: (1) advice aimed at framing the annotation strategy, (2) assistance in identifying and resolving differences of opinion between annotators, (3) rentability indicators for each expert intervention, and (4) methods for analyzing the relevance of the learning base under construction.

In conclusion, this thesis provides an innovative approach to design a learning base for a conversational assistant, involving domain experts for their actual knowledge, while requiring a minimum of analytical and technical skills. This work opens the way for more accessible methods for building such assistants.

# 2. ABSTRACT VULGARISÉ (en moins de 1000 caractères)

## 2.A. En français

La tâche d'annotation, nécessaire à l'entraînement d'assistants conversationnels, fait habituellement appel à des experts du domaine qui est à décrire. Néanmoins, sa complexité nécessite de solides compétences analytiques. De ce fait, la plupart des projets d'annotation choisissent de former les experts aux tâches d'analyse pour en faire des "super-experts".

Dans cette thèse, nous avons plutôt décidé de valoriser les connaissances réelles des experts en proposant une nouvelle méthode d'annotation. Celle-ci se base sur une coopération Homme/Machine, où l'expert caractérise les données deux à deux en fonction de la similarité de leur cas d'usage, et où la machine utilise ces associations pour proposer un jeu d'entraînement pertinent. Nous avons montré que cette approche réduit sensiblement la complexité de la tâche tout en réduisant la nécessité de formation, permettant ainsi aux experts d'intervenir plus facilement dans un projet d'annotation.

## 2.B. In English

Usually, the annotation task, used to train conversational assistants, relies on domain experts to describe it. Nevertheless, its complexity requires strong analytical skills. As a result, most annotation projects choose to train experts in analytical tasks to turn them into "super-experts".

In this thesis, we have chosen to value the real knowledge of experts by proposing a new annotation method. This method is based on a Human-Machine cooperation, where the expert describes the data pairwise based on the use case similarity, and the machine uses these associations to suggest a relevant training dataset. We demonstrated that this approach significantly reduces the task's complexity while decreasing the need for training, making it easier for experts to participate in an annotation project.