Thèse Erwan SCHILD

Plan détaillé

[Remerciements 2](#_Toc124348242)

[I. Introduction : « Concevons un robot conversationnel ! » 2](#_Toc124348243)

[II. Etat de l’art : *Comment fait-on un chatbot ?* 2](#_Toc124348244)

[II.A. Les différents types de chatbot 2](#_Toc124348245)

[II.B. Organisation d’un projet de création d’un chatbot 3](#_Toc124348246)

[II.B.1. Définition des acteurs 3](#_Toc124348247)

[II.B.2. Chronologie d’un projet 3](#_Toc124348248)

[II.B.3. Complexité de la collecte de données et Biais éventuels 3](#_Toc124348249)

[II.C. Création d’une base d’apprentissage pour notre chatbot 4](#_Toc124348250)

[II.C.1. Création « manuelle » de la d’apprentissage 4](#_Toc124348251)

[II.C.2. Création assistée par des regroupements non-supervisés 4](#_Toc124348252)

[II.C.3. Conception assistée par des regroupements semi-supervisés 4](#_Toc124348253)

[II.C.4. Conception basée sur des méthodes d’apprentissage actif 5](#_Toc124348254)

[II.D. Quelques connaissances techniques (***#ANNEXE ?***) 5](#_Toc124348255)

[III.D.1. Les algorithmes de clustering 5](#_Toc124348256)

[III.D.2. Evaluation d’une clustering 5](#_Toc124348257)

[III. Méthode proposée : Le *clustering interactif* 5](#_Toc124348258)

[III.A. Description théorique de la méthode 5](#_Toc124348259)

[III.B. Espoirs de la méthode proposée 6](#_Toc124348260)

[III.C. Description technique et implémentation 6](#_Toc124348261)

[III.D. Protocole d’utilisation : Mode d‘emploi associé (***#CONCLUSION ?***) 6](#_Toc124348262)

[IV. Différentes études : Est-ce que cette méthode est viable ? 7](#_Toc124348263)

[IV.A. Etude de **viabilité** : « *est-ce que la méthode marche/converge ?* » 7](#_Toc124348264)

[IV.B. Etude **technique** : « *quelle est la meilleure implémentation ?* » 7](#_Toc124348265)

[IV.C. Etude des **coûts** : « *quels sont les coûts à investir ?* » 8](#_Toc124348266)

[IV.D. Etude des **erreurs** : « *quel est l’impact d’une différence d’annotation ?* » 8](#_Toc124348267)

[IV.E. Etude **métier** : « *comment interpréter les résultats et leur donner du sens ?* » 9](#_Toc124348268)

[IV.F. Etude d’**arrêt** : « *quand le résultat est-il satisfaisant ? »* 9](#_Toc124348269)

[IV.G. **Autres** études à réaliser 10](#_Toc124348270)

[V. Conclusions (#TODO) 10](#_Toc124348271)

[V.A. Rappel de la problématique : comment aider la création d’un jeu de données pour une tâche de classification ? 10](#_Toc124348272)

[V.B. Avantages et limites de la méthode 11](#_Toc124348273)

[V.C. Piste d’ouverture 11](#_Toc124348274)

[Annexes 12](#_Toc124348275)

[A. Les implémentations 12](#_Toc124348276)

[B. Les jeux de données 12](#_Toc124348277)

[C. Les protocoles d’analyses 12](#_Toc124348278)

[D. Complément sur les métriques (homogénéité, complétude, vmeasure) 12](#_Toc124348279)

[E. Complément sur la FMC (Features Maximization) 12](#_Toc124348280)

[Liste des algorithmes 12](#_Toc124348281)

[Liste des figures 12](#_Toc124348282)

[Liste des tableaux 12](#_Toc124348283)

[Bibliographie 12](#_Toc124348284)

Abstract / Résumé

|  |
| --- |
| (à faire à la fin) |

# Remerciements

|  |
| --- |
| (à faire à la fin)   * Merci à ChatGPT d’avoir générer une partie de mon manuscrit :p |

# I. Introduction : « Concevons un robot conversationnel ! »

|  |
| --- |
| (à faire à la fin)   * **Accroche** : Relation client à distance, accès base documentaire, … * **Chatbot** : Essor de leur utilisation, Révolution technique fréquente, …   + Avec moteur de règles   + Avec classification d’intention   + Avec modèles de langue et génération * **Besoin de données**:   + Collecte de données spécifiques   + Problèmes de biais, de RGPD |

# II. Etat de l’art : *Comment fait-on un chatbot ?*

|  |
| --- |
| **Transition / Récap** : |

## II.A. Les différents types de chatbot

|  |
| --- |
| * **Moteurs de règles**   + Basé sur la détecté de mots clés   + + : facile à mettre en œuvre   + - : peu robuste au langage naturel   + Paramétrage des réponses * **Paramétrage intentions-entités**   + Classification d’intention et/ou détection d’entités   + + : plus robuste au langage naturel, facile à paramétrer, réponses contrôlées   + - : demande de l’entrainement, des données, …   + Paramétrage des réponses * **Génération de réponse**   + Réseau de neurones avec attention   + Transformers   + + : plus robuste   + - : plus complexe à mettre en œuvre, réponses non contrôlées   + Réponses non paramétrées * **Approche hybride :**   + Cumul des trois approches pour cumuler certains avantages suivant les besoins |
| **Transition / Récap** : On se concentre sur le paramétrage intention-entités ici car c’est le plus commun à date de cette étude |

## II.B. Organisation d’un projet de création d’un chatbot

### II.B.1. Définition des acteurs

|  |
| --- |
| * **Data scientistes :**   + Experts en IA   + Peu de connaissance métier, i.e. peu de regarde critique sur la pertinence des résultats (autre que statistique) * **Expert métier :**   + Pas de connaissance en IA, i.e. nécessitent des formations   + Connaissance métier forte, i.e. peuvent décrire la pertinence d’un résultat * **Chef de projet**   + Pas de connaissance en IA   + Pas de connaissance métier   + Connaissance du besoin (hypothèse non vérifiée car parfois ils ne savent pas ce qu’ils veulent dû à la méconnaissance des capacités de l’IA) |

### II.B.2. Chronologie d’un projet

|  |
| --- |
| * **Cadrage du projet** * **Collecte des données** * **Modélisation d’une structure et Labellisation des données** * **Entrainement et Test** * **Déploiement de la première version** * **Annotation du comportement, Amélioration et Déploiement des versions suivantes** |

### II.B.3. Complexité de la collecte de données et Biais éventuels

|  |
| --- |
| * **Souvent pas de données à disposition :**   + En R&D, « 80% » sur la recherche d’algo sur des données **publiques**   + En entreprise, « 80% » sur la gestion des données **privées/spécifiques** sur des algo connus * **Risque de biais dans les données :**   + Biais **d’échantillon** : la collecte ne représente pas la réalité   + Biais de **sélection** : le trie de la collecte ne représente plus la réalité   + Biais de **confirmation** : on garde les données qui nous arrangent   + Biais de **valeur** : les données ne sont pas éthiquement représentatives   + Biais de **contexte** : les données d’un cas d’usage ne sont pas toujours réutilisables pour un autre cas d’usage (ex : différence entre les jargons des AV clients et celui des AV conseillers) |

## II.C. Création d’une base d’apprentissage pour notre chatbot

### II.C.1. Création « manuelle » de la d’apprentissage

|  |
| --- |
| * **Chronologie classique**   + Définition d’une structure en atelier et Annotation des données   + Premier conflit : La structure est trop théorique   + Redéfinition et Ré-annotation   + Second conflit : Les structure ou les données ne sont pas adaptées   + Collecte complémentaire, Redéfinition et Ré-annotation * **Avantages :**   + Transmission progressive du savoir aux datascientist   + Test des modélisations potentielles * **Inconvénients :**   + Nombreux ateliers   + Nombreuses remises en questions / aller-retour de conception   + Se base sur de la connaissance que les experts métiers n’ont pas   + Comment les aider dans ce problème d’organisation ? |

### II.C.2. Création assistée par des regroupements non-supervisés

|  |
| --- |
| * **Constat :**   + Pour des jeux de données à taille humaine (moins de 20.000 données), le premier tri est manuel sur la base des patterns commun (ordonnancement alphabétique)   + *Rappel phase initialisation analyse métier H1* * **Solution :**   + Un clustering pourrait simplifier cette tâche ! * **Avantages :**   + Regroupement automatique   + Découverte de la structure * **Inconvénients** :   + Les résultats sont souvent peu pertinents   + Similarité par entités, et pas par intentions   + Nuances métiers non comprises   + Plusieurs soucis si le jeu de données est déséquilibré ou spécifique   + Absence d’un modèle de langue spécifique au contexte… |

### II.C.3. Conception assistée par des regroupements semi-supervisés

|  |
| --- |
| * **Solution :**   + On peut envisager ainsi de corriger le clustering en y insérant des contraintes métiers   + Méthodes semi-supervisée * **Avantage :**   + On a réglé les problèmes de pertinence en ajoutant des contraintes * **Inconvénients** :   + Choisir comment modéliser ces contraintes peut être complexe   + Surtout énorme en ajoutant des contraintes   + Choisir les contraintes pertinentes est une tâche difficile |

### II.C.4. Conception basée sur des méthodes d’apprentissage actif

|  |
| --- |
| * **Solution :**   + On peut demander à la machine de définir les contraintes dont elle a besoin pour s’améliorer / confirmer son comportement   + On peut séparer et cibler les tâches pour que le clustering se nourrissent des commentaires de l’expert et que l’expert corrige ce qui semble utile au clustering   + Sous-entendu : Préférer la collaboration à la supériorité (que ce soit celle de la machine ou celle de l’expert) * **Avantage :**   + On a réglé les problèmes de pertinence et de coûts en ajoutant des contraintes * **Inconvénients / problème à résoudre :**    + Accepter de collaborer avec la machine (problème UX, ergo, accompagnement au changement)   + Il faut prouver cette méthode |

## II.D. Quelques connaissances techniques (***#ANNEXE ?***)

### III.D.1. Les algorithmes de clustering

|  |
| --- |
| * **Kmeans** * **Hierarchique** * **Spectral** * **DBScan** * **Affinity Propagation** |

### III.D.2. Les interactions avec le clustering

|  |
| --- |
| * **Sur les données / sur le résultat :** ajouts de contraintes sur les données, suppressions ou modifications manuelles de données, réorganisation manuelles des clusters, … * **Sur les paramètres :** modifier les hyper-paramètres, modifier le nombre de clusters, modifier les embeddings, utiliser d’autres algorithmes, … * **Sur les propositions de la machine :** détection de données aberrantes, suggestion de modifications, … * **Besoin de visualisation :** vue des contraintes, de la représentation vectorielle, … |

### 

### III.D.3. Evaluation d’une clustering

|  |
| --- |
| * **Homogénéité – Complétude – Vmeasure** * **FMC** |

# III. Notre proposition de *clustering interactif*

|  |
| --- |
| **Transition / Récap** : On a vu que :   * Le travail est principalement manuel : **comment peut-on l’assister ?** * La définition de la structure de classes est un mélange de connaissances métiers et de regroupement manuel sur la base de patterns linguistiques commun : **faisons du clustering !** * Mais le clustering est souvent peu pertinent pour un usage métier : **intégrons-y des contraintes !** * L’ajout de contraintes métiers est fastidieux : **ajoutons de l’active learning !** |

## III.A. Description théorique de la méthode

|  |
| --- |
| **Nous proposons de combiner les techniques vues précédemment :**   * **Clustering sous contraintes**   + Kmeans : Classique, Incontournable, Rapide, Efficace   + Hiérarchique : Lent mais facile à implémenter   + Spectral : Permet des topologies complexes   + DBScan : Classique, Incontournable, Rapide, Efficace, Peu d’hyperparamètres   + Affinity propagation :   + Metric learning : Lent mais plus adapté au corpus   + … * **Echantillonnage de contraintes à annoter**    + Random ou Pseudo-random   + Farhtest : Scinder les gros clusters   + Closest : Redéfinir la position des frontières de clusters   + … * **Annotation de contraintes**   + MUST-LINK / CANNOT-LINK / SKIP   + « Répondriez-vous de la même manière à ces deux demandes ? » * **Boucle itérative entre clustering, échantillonnage et annotation**   + Améliorer le résultat précédent   + Autant de boucle que « nécessaire »   + Avoir le clustering le plus efficace pour avoir de bon résultats   + Avoir l’échantillonnage le plus efficace pour améliorer le plus efficacement   + Avoir une annotation sans ambiguïté pour ne pas biaiser la construction itérative * **Analyses**   + Analyse de l’évolution de l’accord clustering->annotation   + Analyse des patterns linguistiques pertinents   + Analyse de la formation de clusters (taille, répartition, …) * NB : Réutilisation de schéma etc |

## III.B. Espoirs de la méthode proposée

|  |
| --- |
| * Moins de formations, d’ateliers, … * Se concentrer sur son domaine de compétence (i.e. pas de datascience pour les experts métiers) * Permettre de trouver la base d’apprentissage * Méthode réaliste / pas trop coûteuse * … |

## III.C. Description technique et implémentation

|  |
| --- |
| * cognitivefactory.interactive-clustering : Gestion des données * cognitivefactory.interactive-clustering : Gestion des contraintes + conflits * cognitivefactory.interactive-clustering : Algorithmes de clustering * cognitivefactory.interactive-clustering : Algorithmes de sampling * cognitivefactory.interactive-clustering-gui : Interface d’annotation * cognitivefactory.interactive-clustering-gui : Interface d’analyse * NB : captures d’écrans pour donner un aperçu, puis redirection vers les annexes |

## III.D. Protocole d’utilisation : Mode d‘emploi associé (***#CONCLUSION ?***)

|  |
| --- |
| * Collecte des données * Itération de clustering > échantillonnage > annotation   + A chaque conflit : correction nécessaire   + A la fin d’un clustering : caractériser le résultat avec FMC   + A chaque itération : voir l’évolution par rapport à la précédente |

# IV. Différentes études : Est-ce que cette méthode est viable ?

|  |
| --- |
| **Transition / Récap :** Cette méthode étant nouvelle, plusieurs études sont nécessaires :   1. **Est-ce que la méthode marche/converge ?** 2. **Quelle est la meilleure implémentation ?** 3. **Quels sont les coûts à investir ?** 4. **Quel est l’impact d’une différence / erreur d’annotation ?** 5. **Comment interpréter les résultats et leur donner du sens ?** 6. **Quand s’arrêter ?**   (à formuler sous la forme d’hypothèse de travail à vérifier ?  et/ou formuler/rappeler ces hypothèse en début de chaque sous-partie) |
| * **Remarque en préambule** : C’est compliqué de remettre en cause l’annotation manuelle avec des vérité terrain conçues manuellement… Certaines études ne sont donc pas faisables |

## IV.A. Etude de **viabilité** : « *est-ce que la méthode marche/converge ?* »

|  |
| --- |
| **Etude de convergence vers une vérité terrain établie :**   * **Expérience** : Faire des itérations de « clustering > échantillonnage > annotation » jusqu’à convergence vers la vérité terrain   + Jeu de données : Carte bancaire (500 questions)   + Cf. EGC/IJDWM * **Résultats :**   + Les contraintes sont respectées   + La vérité terrain est atteinte (La vmeasure atteint 100%)   + L’ajout de contraintes permet vite d’être meilleur qu’un clustering simple * **Conclusion** : la méthode est viable : on trouve bien vérité terrain   + **Avantage :** Pas besoin de définir une structure, elle se découvre toute seule   + **Avantage :** Besoin de peu de connaissance en IA (same ? oui ou non)   + **Inconvénient**: Besoin de « beaucoup » de contraintes   + 🡺 **IL FAUT UNE OPTIMISATION (IV.B.)** * **Discussion** : balance entre annotations de contraintes et clustering   + Si pas assez de contraintes : alors clustering pas assez pertinent !   + Si 100% de contraintes : alors résultat trop subjectif !   + Trouvons une juste milieu : supposons 80% pour la suite (annotation partielle) |

## IV.B. Etude **technique** : « *quelle est la meilleure implémentation ?* »

|  |
| --- |
| **Etude des paramètres optimaux :**   * **Expérience** : Faire des itérations de « clustering > échantillonnage > annotation » jusqu’à convergence vers la vérité terrain et étudier des itérations clés pour déterminer quelle implémentation est la plus efficace.   + Jeu de données : Carte bancaire (500 questions)   + Paramètres étudiés : prétraitement, vectorisation, sampling, clustering   + Métrique : VMeasure du clustering obtenue avec la vérité terrain en fonction du nombre de contraintes annotées   + Cf. EGC/IJDWM * **Résultats :**   + Tous les paramètres sont importants, surtout le sampling et le clustering   + Meilleure implémentation trouvée * **Conclusion** :   + Il y a donc moyen d’optimiser la méthode pour annoter moins de contraintes * **Remarques**:   + Expliquer le choix du kmeans : rapide, efficace, itération rapide   + Expliquer le choix du closest : favorise les MUST-LINK   + Expliquer le non-choix du farthest : favorise CANNOT\_LINK, mais n’aide pas |

## IV.C. Etude des **coûts** : « *quels sont les coûts à investir ?* »

|  |
| --- |
| **Etude du temps d’annotation**   * **Expérience** : Faire annoter des contraintes par plusieurs annotateurs   + Jeu de données : Carte bancaire (500 questions) ou Titre de journaux (**???**) * **Résultats** : * **Conclusions** : * **Discussions** : Définition d’un batch moyen   --------------------------------------------------------------------------------------------------------------------------  **Etude du temps de calcul**   * **Expérience**: Estimer le temps de calcul de chaque algorithme   + Jeu de données : Carte bancaire (1000 questions, artificiellement augmenté jusqu’à 1000 données)   + Estimation des facteurs influents : nombre de données, nombre de contraintes, nombre de clusters, hyper-paramètres, effets aléatoires, … * **Résultats** :   + Estimation des temps de calcul et des facteurs influents * **Conclusions** :   + Certains algorithmes sont longs…   + Définition d’une fonction d’estimation du temps de calcul pour chaque algo * **Discussion** : balance entre performance et coût temporel   + Les itérations doivent être fluides   + On a plus à gagner à ajouter des contraintes qu’attendre un algo trop long   + Un optimal serait d’annoter des batch d’une durée d’un algo pour ne pas avoir trop à attendre entre deux itérations |

## IV.D. Etude des **erreurs** : « *quel est l’impact d’une différence d’annotation ?* »

|  |
| --- |
| **Préambules**   * **Source potentielle des différences** :   + Différence de points de vue   + Erreurs d’inattention   + Données ambigües * **Identification des différences** :   + Relectures / revue d’annotation   + Détection des incohérences entre contraintes   + Adjudication / ajout de redondance   --------------------------------------------------------------------------------------------------------------------------  **Etude de l’impact d’une erreur par simulation**   * **Expérience** : Faire des itérations de « clustering > échantillonnage > annotation » en simulant un pourcentage d’erreur d’annotation   + Jeu de données : Carte bancaire (500 questions)   + Pourcentage d’erreur variable   + Prise en compte ou non de la résolution de conflits * **Résultats** :   + Peut gravement impacter le résultat   + N’est pas détectable sans redondance * **Conclusions** : * **Discussions** : importance de la fiabilité des annotations   + Besoin de savoir ce que l’on veut : Ajouter de l’adjudication en début de projet pour confronter rapidement les visions   + Besoin de limiter les erreurs : Ajout de redondance ou de stratégie de vérification (de nouvelles méthode de sélection) mais cela entraîne un surcoût   + Préférer passer l’annotation que de forcer une annotation ambiguë   --------------------------------------------------------------------------------------------------------------------------  **Etude de ré-annotation d’un chatbot existant**   * **Expérience** : Ré-annoter un AV existant pour identifier les inconvénients pratiques   + Jeu de données : Moyens de paiements (8000 questions de production)   + Deux annotateurs avec des règles « floues » (comme c’est le cas en initialisation) mais avec un « objectif » commun (arbre de dialogues & réponses « connues »)   + Revues d’annotations à mi-projet * **Résultats** :   + 23% de désaccord, mais au moins 77% d’accord sans concertation !   + Après revue d’annotation : (**???**)   + **Expériences interrompue** * **Conclusions** : * **Discussions** : Quelques conseils   + En début de projet : Adjudication pour confirmer les visions   + Correction des incohérences le plus rapidement possible   + D’autres sélection pour ajouter de la redondance   + Mais cela représente un cout supplémentaire |

## IV.E. Etude **métier** : « *comment interpréter les résultats et leur donner du sens ?* »

|  |
| --- |
| **Préambule de définition de la FMC ou référence à l’état de l’art / annexe**  **Etude de la caractérisation du clustering par FMC**   * **Expérience** : Faire des itérations de « clustering > échantillonnage > annotation » jusqu’à convergence vers la vérité terrain et étudier la FMC du clustering en cours   + Jeu de données : Carte bancaire (500 questions)   + Métrique : VMeasure adaptée pour comparer la FMC d’un clustering et la FMC de la vérité terrain * **Résultats** : * **Conclusions** : * **Discussions** : |

## IV.F. Etude d’**arrêt** : « *quand le résultat est-il satisfaisant ? »*

|  |
| --- |
| **Etude d’un prérequis d’arrêt : la cohérence du clustering obtenu**   * **Expérience** : Faire des itérations de « clustering > échantillonnage > annotation » jusqu’à convergence vers la vérité terrain et étudier si un classifier entrainé sur cette base est stable, i.e. s’il arrive à retrouver sa base d’apprentissage   + Jeu de données : Carte bancaire (500 questions)   + Remarque : si la base d’apprentissage n’est pas stable, le clustering doit encore être modifié… * **Résultats** : * **Conclusions** : * **Discussions** :   --------------------------------------------------------------------------------------------------------------------------  **Etude d’un critère d’arrêt : l’accord entre un batch d’annotation et le clustering précédent**   * **Expérience** : Faire des itérations de « clustering > échantillonnage > annotation » jusqu’à convergence vers la vérité terrain et étudier l’évolution des accords/désaccords entre l’annotateur et le clustering   + Jeu de données : Carte bancaire (500 questions)   + Remarque : si l’accord est maximal, l’annotateur n’a plus de valeur ajoutée car le clustering n’est jamais modifiée * **Résultats** : * **Conclusions** : * **Discussions** :   --------------------------------------------------------------------------------------------------------------------------  **Etude d’un critère d’arrêt : la similitude entre deux clustering consécutifs**   * **Expérience** : Faire des itérations de « clustering > échantillonnage > annotation » jusqu’à convergence vers la vérité terrain et étudier l’évolution des similitudes entre le clustering courant et le clustering précédent   + Jeu de données : Carte bancaire (500 questions)   + Remarque : si la similitude est forte, c’est que le clustering devient stable * **Résultats** : * **Conclusions** : * **Discussions** :   --------------------------------------------------------------------------------------------------------------------------  **Etude d’un critère d’arrêt : la similitude entre deux FMC de clustering consécutifs**   * **Expérience** : Faire des itérations de « clustering > échantillonnage > annotation » jusqu’à convergence vers la vérité terrain et étudier l’évolution des similitudes entre la FMC du clustering courant et la FMC du clustering précédent   + Jeu de données : Carte bancaire (500 questions)   + Remarque : si la similitude est forte, c’est que le clustering devient stable * **Résultats** : * **Conclusions** : * **Discussions** : |

## IV.G. **Autres** études à réaliser

|  |
| --- |
| 1. **Choix du nombre de clusters** 🡪 problème de recherche complexe    * **Piste de résolution** : plusieurs clusterings + vote collaboratif ? algorithmes sans le nombre de clusters en hyper-paramètres 2. **Impact d’un modèle de langage** 🡪 nécessite de nombreuses données spécifiques au domaine    * **Piste de résolution** : script d’étude comparative déjà prêt, mais il manque les données opensources… 3. **Paradigme d’annotation** (intention vs dialogue) 🡪 problème d’UX + objectif métier    * Etude Ergo, sort de mon domaine d’expertise 4. (et plein d’autres que j’ajouterai au fur et à mesure de ma rédaction) |

# V. Conclusions (**#TODO**)

## V.A. Rappel de la problématique : comment aider la création d’un jeu de données pour une tâche de classification ?

|  |
| --- |
|  |

## V.B. Avantages et limites de la méthode

|  |
| --- |
|  |

## V.C. Piste d’ouverture

|  |
| --- |
|  |

# Annexes

## A. Les implémentations

## B. Les jeux de données

## C. Les protocoles d’analyses

## D. Complément sur les métriques (homogénéité, complétude, vmeasure)

## E. Complément sur la FMC (Features Maximization)

# Liste des algorithmes

# Liste des figures

# Liste des tableaux

# Bibliographie