Intérêt de l'enrichissement sémantique pour une tâche de catégorisation de textes courts par méthode hybride avec peu de données d'entraînement

Armelle Ramond

Intérêt de l'enrichissement sémantique pour une tâche de catégorisation de textes courts par méthode hybride avec peu de données d'entraînement

Ramond
Armelle

Sous la direction de Georges Antoniadis

Réalisé au sein de la société : Holmes,

Sous la supervision de Luca Dini et Muntsa Padró

UFR LLASIC
Département I3L

Mémoire de master 2 recherche - 30 crédits - Mention Sciences du langage

Spécialité : Industries de la langue

Parcours : Traitement automatique de la langue écrite et de la parole

Année universitaire 2015-2016

Intérêt de l'enrichissement sémantique pour une tâche de catégorisation de textes courts par méthode hybride avec peu de données d'entraînement

Ramond
Armelle

Sous la direction de Georges Antoniadis

Réalisé au sein de la société : Holmes,

Sous la supervision de Luca Dini et Muntsa Padró

UFR LLASIC
Département I3L

Mémoire de master 2 recherche - 30 crédits - Mention Sciences du langage

Spécialité : Industries de la langue

Parcours : Traitement automatique de la langue écrite et de la parole

Année universitaire 2015-2016

Remerciements

Tout d'abord, je tiens à remercier Luca qui m'a proposé ce sujet de travail qui correspondait tout à fait à mes intérêts. Merci pour ses conseils avisés et la confiance qu'il m'a accordée et continue de m'accorder pour poursuivre ce projet et d'autres encore.

Je remercie M. Antoniadis pour son encadrement et son suivi au cours de ce travail de mémoire.

Un très grand merci également à Muntsa qui a attrapé ce projet au vol et m'a accompagnée dessus quotidiennement. Son encadrement, ses conseils judicieux, et son implication ont donné une autre dimension à ce travail. Je lui suis particulièrement reconnaissante pour notre partenariat très enrichissant.

Je remercie également Mathieu pour sa pédagogie, sa patience et sa bonne humeur constante. Merci de m'avoir guidée dans les méandres de Java et du SVN.

Enfin, un merci à toute l'équipe Holmes pour l'accueil et l'ambiance, et notamment à David pour ses propositions de titres très pertinentes.

UNIVERSITÉ Grenoble Alpes

Déclaration anti-plagiat
Document **à scanner** après signature
et **à intégrer** au mémoire électronique

DECLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : RAMOND PRENOM : ARMELLE

DATE : 23/06/2016 SIGNATURE [redacted]

Mise à jour avril 2016

Sommaire

Introduction

A Contexte et sujet de travail

Le travail réalisé dans le cadre de ce mémoire s'inscrit dans un contexte de recherche industrielle. Il répond à un besoin émergent de la collaboration de deux entreprises : Holmes Semantic Solutions, spécialiste de l'analyse sémantique, et Eloquant, qui propose des services dans le domaine de la relation client. L'objectif est de développer un système d'analyse de verbatims clients. Dans le domaine de la relation client, on entend par "verbatim" la réponse textuelle d'un client à une question ouverte dans le cadre d'une enquête. Ce système se décline en analyse d'opinion, détection de sentiments et classification.

En terme de classification automatique, Holmes dispose déjà de systèmes développés pour des contextes spécifiques, mais difficilement transposables à de nouveaux domaines d'activité. Il s'agit donc de mettre au point un système plus générique qui couvre le secteur de la relation client et qui soit en mesure de catégoriser automatiquement des verbatims clients. Or, comme le montre l'état de l'art, il n'existe pas aujourd'hui de solution opérationnelle répondant à un tel besoin.

Ce mémoire décrit en trois partie notre travail de recherche sur l’intérêt de l’enrichissement sémantique pour une tâche de classification. Nous présentons dans un premier temps les données que nous allons exploiter ainsi que leur structuration. Puis, nous expliquons notre méthodologie de développement du système, avant d’explorer en détail l’évaluation mise en œuvre et les résultats obtenus.

B Etat de l'art

1. *La catégorisation automatique*

La classification automatique de texte est une tâche qui a pour objectif d'automatiser l'attribution d'une ou plusieurs classes à un document en fonction de son contenu. Lorsque les classes ne sont pas définies à l'avance mais élaborées au cours de la tâche de classification, on parle de "clustering", et l'on est dans le cadre d'une "classification non supervisée". A l'inverse, la "catégorisation" permet de répartir les documents dans des classes fixées en amont de la tâche. C'est cette deuxième méthode, la "classification supervisée", qui nous intéresse.

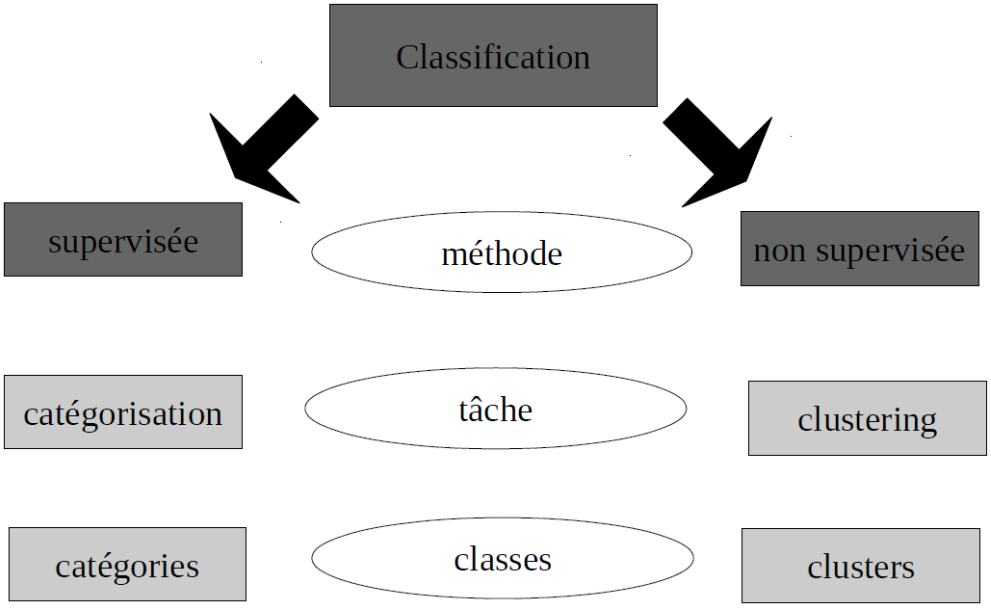


Figure 1 : Méthodes, tâches et classes en classification automatique

L'activité dans laquelle s'insère notre travail est celle d'une classification de verbatims clients dans des catégories déterminées à l'avance par le commanditaire de la tâche (il s'agit ici d'opérateurs de téléphonie et internet). Les spécificités des données à traiter doivent être prises en compte car elles impactent fortement les méthodes de traitement et les résultats. En effet, nous sommes face à des documents textuels particulièrement courts, parfois de quelques mots seulement, contenant généralement une à deux phrases seulement. D'autre part, la nature de ces textes et leur mode de communication (sms, web, chat, e-mail, etc.) engendrent des contenus peu standards d'un point de vue linguistique : orthographe, syntaxe, ponctuation ne répondent pas aux règles conventionnelles de la langue. Enfin, nous ne disposons que de très peu de données déjà catégorisées ce qui est un frein pour les techniques d'apprentissage automatique.

2. *Méthodologie de recherche bibliographique*

L'étude de la littérature scientifique correspondant à notre sujet de recherche s'est faite selon une méthodologie en deux étapes. Dans un premier temps a été réalisée une exploration assez large de la documentation relative aux systèmes de classification automatique de texte. Cela a permis d'en appréhender les grands principes ainsi que les méthodes et les outils reconnus comme performants. Dans un deuxième temps, cette recherche a été approfondie et précisée en se focalisant sur les contraintes spécifiques à notre travail (textes courts, petit corpus d'entraînement) et notre approche de spécialité (le TAL, et plus précisément la sémantique). De plus, les références les plus pertinentes citées dans la documentation explorée au cours de la première étape ont été recherchées.

Ainsi, un panorama des techniques de classification automatique a été dressé. La littérature est particulièrement riche concernant les techniques d'apprentissage automatique, beaucoup moins pour les solutions exploitant le niveau sémantique. D'autre part, il faut noter que les publications explorées, bien qu'en lien avec notre sujet, peuvent s'avérer plus larges ou parallèles à notre domaine : c'est le cas de la classification de données par rapport à la classification de textes, ou encore la fouille de texte par rapport à la classification de texte. Enfin, il faut prêter une attention particulière aux caractéristiques de données, telles que la taille des textes et du corpus d'entraînement : selon les auteurs, les dimensions ne sont pas du tout du même ordre.

3. *Classification de textes : principes et outils*

a Deux approches : symbolique et statistique

L'état de l'art montre que la classification automatique de textes relève de deux familles de méthodes : symboliques et statistiques.

Les méthodes symboliques se basent sur des règles de grammaire définies par des experts, en l'occurrence des linguistes. Le principe est le suivant : le texte est soumis à une règle et en fonction du retour (si le texte répond à cette règle ou non), on peut attribuer une ou plusieurs classes, ou du moins faire varier sa probabilité d'appartenance aux différentes classes. En terme de résultats, la précision est bonne mais le rappel faible. Ces méthodes présentant un coût important, elles ne sont mises en œuvre que dans les cas où les autres méthodes ne s'avèrent pas suffisamment performantes, par exemple lorsque les corpus sont petits. D'autre part, elles présentent l'avantage d'avoir une lisibilité accrue dans

l'interprétation des résultats et sont donc pertinentes dans une optique d'amélioration des méthodes.

Les méthodes statistiques consistent en la représentation du texte analysé selon des caractéristiques numériques déterminées, comme par exemple la fréquence d'un mot ou d'un lemme. Chaque caractéristique, appelée "trait", représente une dimension dans un espace. On peut ainsi utiliser cet espace multidimensionnel pour situer les documents et les classes les uns par rapport aux autres. Ce sont ces méthodes qui sont largement décrites dans la littérature car elles sont performantes et donc couramment utilisées.

De ces deux familles, découle une troisième : les méthodes hybrides combinent les deux approches citées précédemment.

b L'apprentissage automatique

La mise en œuvre de la classification automatique de texte repose à l'heure actuelle principalement sur l'apprentissage automatique. Selon cette technique, une machine est en mesure d'apprendre de façon systématique les critères et règles à appliquer par la suite aux nouvelles données en entrée (dans notre cas, des textes) afin de leur attribuer une ou plusieurs classes. De même que nous l'avons vu pour la classification, lorsque les classes sont prédéterminées, on parle d'"apprentissage supervisé". Celui-ci peut être dit "probabiliste" si la sortie consiste en l'attribution d'une probabilité d'appartenance à une classe.

L'apprentissage automatique pour une tâche de classification supervisée ou catégorisation de texte comporte deux phases. La première est appelée "phase d’entraînement". Elle requiert un "corpus d’entraînement", c’est-à-dire un ensemble de données déjà étiquetées (on a attribué manuellement des catégories aux textes). Un algorithme va alors étudier ces données selon une grille de traits choisis afin de déterminer un modèle de classement. Ensuite, lors de la "phase de test", l'algorithme utilisera le modèle construit pour attribuer automatiquement des étiquettes aux nouvelles données, appelées "données de test".

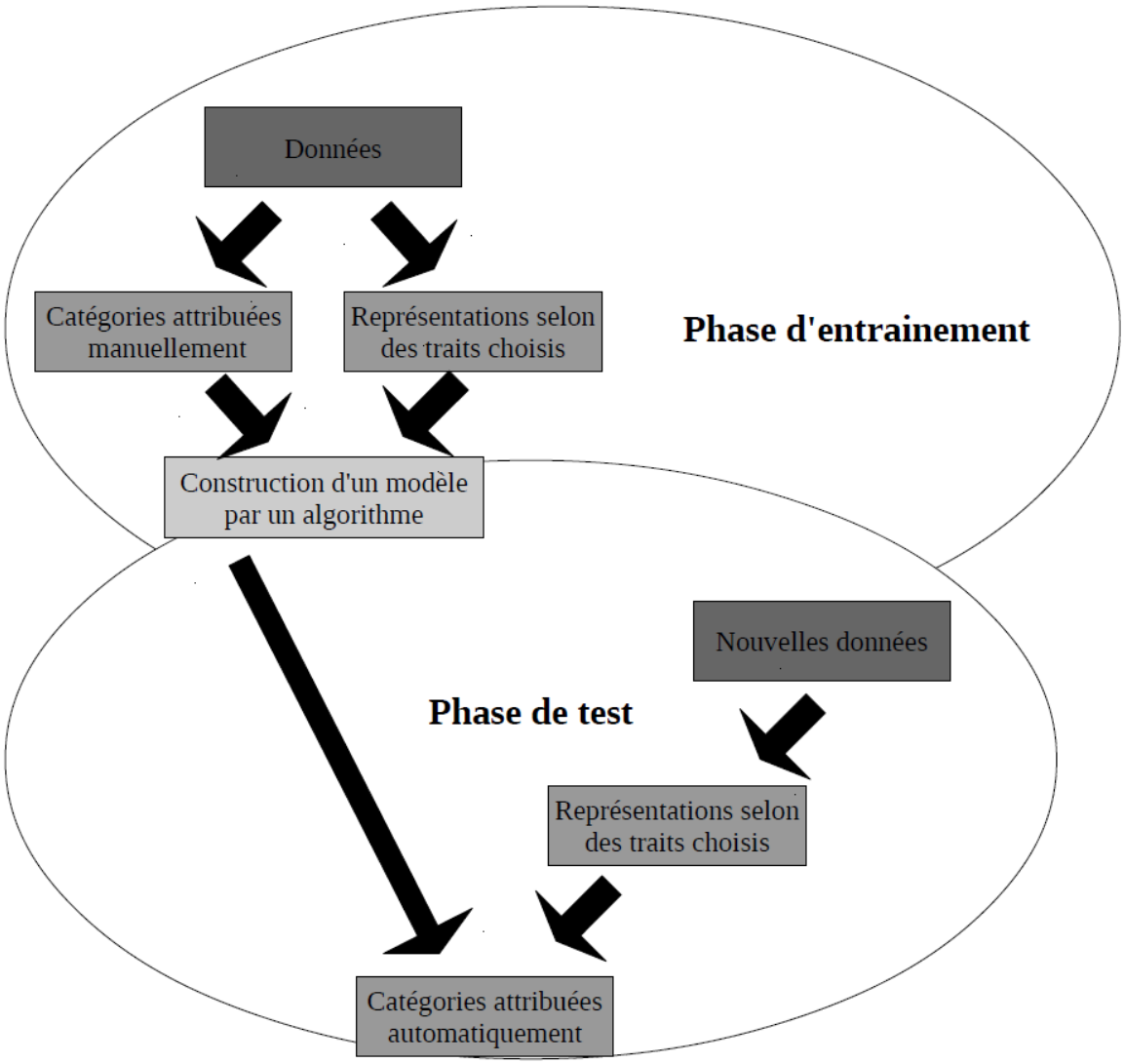


Figure 2 : Apprentissage automatique

[Baharudin, Lee, et Khan, 2010] présente les principes de l'apprentissage automatique pour la classification de textes et dresse un panorama des algorithmes utilisés. L'article met notamment en avant les leviers sur lesquels agir pour obtenir les meilleurs résultats. Cette publication, dont les grandes lignes sont expliquées ci-dessous, constitue donc un élément important de l'état de l'art puisqu'elle fait un balayage vaste des pratiques actuelles et donne un certain nombre de pistes pour la résolution de difficultés qui sont propres à notre contexte. Ces pistes seront explorées dans la section suivante.

Avant toute chose, pour une tâche de classification automatique, un certain nombre de prétraitements sont nécessaires sur les textes d'entrée : tokenisation, suppression des mots vides, lemmatisation, etc. en fonction de la nature des données à traiter et de la catégorisation visée.

L'étape suivante à mettre en œuvre est celle de la représentation du texte. L'objectif est de décrire son contenu sous une forme compacte et exploitable par un algorithme. Pour cela, on choisit un certain nombre de critères de description, les "traits" ("features" en anglais). Parmi les choix les plus communs, on peut citer : la fréquence des termes (le nombre de dimensions pour la représentation est alors égal au nombre de termes distincts dans le corpus) ou encore les n-grammes de caractères. Les auteurs soulignent alors l'importance d'un travail de réduction de la dimensionnalité. En effet, cela peut fortement améliorer les performances de traitement. Après l'extraction des traits disponibles dans le corpus, il s'agit de sélectionner les plus pertinents, c'est-à-dire ceux qui apportent le plus d'information discriminante au regard de la catégorisation. Une fois les traits sélectionnés, les données d'entraînement et les données de test pourront alors être représentées dans un espace multidimensionnel. De plus, des pistes pour réaliser un enrichissement sémantique sont citées afin d'affiner la représentation des documents.

Une fois ce travail préliminaire réalisé, intervient l'algorithme d'apprentissage automatique qui va construire un modèle de classification à partir des données d'entraînement, puis catégoriser les données de test. Plusieurs types d'algorithmes sont décrits dans [Baharudin, Lee, et Khan, 2010], et que l'on retrouve dans la littérature :

- L'algorithme de Rocchio : chaque catégorie est représentée par un "vecteur prototype" (il s'agit de la moyenne des vecteurs des données d'entraînement classées dans cette catégorie, calculés grâce aux traits dans l'espace multidimensionnel). Une pondération est ensuite appliquée sur chaque donnée de test en fonction de la similarité entre le vecteur de cette donnée et le vecteur prototype de chaque classe. L'avantage de cet algorithme est sa simplicité et sa rapidité. Cependant, les performances constatées ne sont pas parmi les meilleures.
- L'algorithme des k-plus proches voisins : basé ici encore sur la représentation des textes dans un espace vectoriel multidimensionnel, on recherche les k données d'entraînement qui sont les plus proches d'une nouvelle donnée de test. La catégorie du nouveau texte est alors attribuée en fonction de celles des k voisins. La difficulté de cette méthode est de définir la valeur de k la plus pertinente. L'algorithme donne de bons résultats mais implique un temps de traitement assez long, d'autant plus si les données d'entraînement et/ou les traits sont nombreux.

- Les arbres de décision : dans ce cas, les feuilles de l'arbre représentent les catégories et les arcs les combinaisons de traits. Les arbres sont relativement faciles à interpréter car ils représentent la structure logique de la classification. Cependant, ils ne sont pas adaptés si le nombre de traits est élevé, et risquent d'amener à une sur-analyse.
- La classification à base de règles de décision se rapproche des méthodes symboliques vues dans la section précédente. Les données d'entraînement permettent de construire une base de règles qui met en regard les traits avec les catégories. Il n'est pas nécessaire de tester toutes les règles pour attribuer une catégorie. Si le nombre de traits est important, il faut réaliser une implémentation heuristique. On constate cependant dans ce cas une baisse des performances.
- La classification naïve bayésienne considère que la présence d'un trait pour une classe est indépendante de la présence d'autres traits. Elle est facile à mettre en œuvre, efficace et ne requiert qu'une petite quantité de données d'entraînement pour l'estimation des paramètres. Cependant, la probabilité d'appartenance à une classe n'est pas estimée de façon précise. Cela pourrait donc être un problème dans le cas de l'application de méthodes symboliques a posteriori, basées sur la valeur de probabilité d'appartenance à une classe. Cette classification est particulièrement utilisée sur des textes comme les e-mails, les contenus web ou encore les spam.
- Les réseaux de neurones artificiels offrent la possibilité de stocker de nombreux cas et donc de traiter des vecteurs de grande dimensionnalité. Le principe de ce type de réseau est que, à chaque niveau, chaque entrée est connectée à une ou plusieurs sorties avec l'application de règles et l'attribution d'une pondération particulière sur laquelle on peut jouer pour améliorer les performances. Cette architecture a donc la capacité de gérer un nombre élevé de traits et également des données bruyantes et contradictoires. Cependant, le coût computationnel n'est pas négligeable et l'inconvénient souvent cité est que l'on peut difficilement comprendre et justifier les résultats obtenus. La phase d'entraînement peut s'avérer longue, tandis que la phase de test est rapide.

- Les algorithmes génétiques permettent de rechercher les meilleurs traits, selon un principe de survie du plus adapté. Il s'agit d'une méthode stricte et simple et peut améliorer la classification.
- Les machine à vecteurs de support sont basées sur le calcul d'un vecteur déterminé par rapport à une distance entre une frontière (qui délimite deux catégories) et un texte dans l'espace de représentation multidimensionnel. La classification se fait selon un principe 1 contre 1 : pour un texte donné, on procède catégorie par catégorie, en déterminant son appartenance ou non à celle-ci. Ces algorithmes sont parmi les plus répandus et sont reconnus pour leurs bons résultats.
- Enfin, la corrélation floue, qui peut être mise en œuvre sur différents algorithmes décrits précédemment, prend en charge une information floue ou des données incomplètes.

4. Verrous à lever

Les principales techniques de classification automatique ayant été décrites, il s'agit maintenant de chercher précisément dans la littérature quelles réponses peuvent être apportées à nos contraintes spécifiques qui sont : des textes courts en entrée, de quelques mots à quelques phrases (verbatims clients) et un corpus d'entraînement contenant peu de données (7300 verbatims étiquetés environ).

a Textes courts

Différentes publications abordent le traitement de textes courts, certaines s'insérant tout à fait dans notre sujet comme [Doucy et Massoussi, 2012] ou [Huang, Murphey, et Ge, 2013] qui s'intéressent aux verbatims clients. Cependant, il faut bien constater que la notion de "textes courts" est floue et ne correspond pas forcément à notre contexte. Par exemple, [Vernier et al., 2009] s'intéresse à la catégorisation d'évaluations dans des billets de blog. Bien que l'expression "textes courts" soit utilisée, nous ne sommes pas face à des documents de même échelle, les auteurs précisant que les textes ont une moyenne de 700 mots. Dans [Chen, Jin, et Shen, 2011], il est question de résumés automatiques pour résultats de moteur de recherche (snippets), de tweets et de descriptions de produit sur le web.

[Poirier, Fessant, et Tellier, 2010] cherche à effectuer une classification de textes très courts, mais d'un type différent du nôtre (avis web sur des films). De plus, l'objectif visé est particulier : fournir les résultats de catégorisation comme données d'entrée à un système de recommandation automatique. Les besoins ne sont donc pas les mêmes. Enfin, le mode de classification diffère du notre. Ici les auteurs visent une classification axiologique (on cherche à savoir si le texte émet un message positif, négatif ou neutre) ce qui n'est pas notre cas (classification thématique), les critères de catégorisation seront donc éloignés. C'est d'ailleurs le cas pour une majorité de publications, l'analyse de sentiments étant un sujet très demandé : [Vincent et Winterstein, 2013], [Eensoo et Valette, 2014], ou encore [Actes de la 11e Défi Fouille de Texte, 2015] où les participants se sont vus attribuer des tâches de classification dans le cadre d'une analyse de sentiments dans des tweets.

b Petit corpus d'entraînement

Dans bien des cas, les corpus d'entraînement cités dans les publications explorées sont de taille conséquente, ou du moins ne représentent pas un frein pour la tâche à réaliser. Quelques sources tout de même en font mention.

La thèse de [Maciej, 2008] présente un système de classification automatique de news ne nécessitant pas de données d'apprentissage. L'approche adoptée se base sur l'exploitation de ressources externes et un travail sur les entités nommées et l'identification des relations.

[Salperwyck et Lemaire, 2011] fait le point sur l'impact de la taille d'un corpus d'apprentissage. Il s'agit d'un point de vue plus large que le nôtre puisque cet article s'inscrit dans le domaine de la fouille de données en général. Il est tout de même intéressant de noter que les auteurs mettent en avant le peu de littérature sur la question de la taille du corpus d'entraînement et l'intérêt de prendre cela en compte comme critère de performance car il peut influencer fortement les résultats.

D'après [Toussaint, 2011], les méthodes symboliques sont particulièrement pertinentes dans le cas de petits corpus d'apprentissage grâce à leur lisibilité et la possibilité de les enrichir.

c Problématique

Il ressort de notre recherche bibliographique des publications qui couvrent partiellement nos contraintes, mais on constate que la combinaison de nos verrous n'a à

l'heure actuelle pas été traitée. On observe cependant plusieurs pistes, qui portent sur le niveau sémantique, souvent évoquées et qui pourraient répondre à nos contraintes. La problématique de notre travail se définit donc ainsi : en quoi le niveau sémantique peut-il être exploité et enrichi pour améliorer une tâche de classification automatique de textes courts avec peu de données d'entraînement ?

5. Pistes de travail

Les pistes relevées se retrouvent à différents stades de la classification, que nous détaillons ci-après.

a Prétraitements et représentation de texte

La sélection d'attributs se révèle utile pour diminuer les temps de traitements et les performances de calcul requises. En effet, une partie des attributs est non pertinente ou redondante. On est confronté à un espace dimensionnel très vaste qu'il peut être intéressant de réduire pour l'exploiter. Cela permet également une meilleure lisibilité du processus d'apprentissage. Il existe différentes méthodes de sélection d'attributs, ainsi que de nombreuses mesures d'évaluation.

Les prétraitements proposés par [Poirier, Fessant, et Tellier, 2010] portent sur la définition de l'espace vectoriel : suppression des termes peu fréquents ou vides de sens, correction et lemmatisation, représentation fréquentielle, découpage en unigrammes (résultats meilleurs qu'en trigrammes). Il est cependant précisé qu'il faut limiter les prétraitements car on observe une baisse des performances si l'on en réalise trop. Ce phénomène est également évoqué dans [Abdaoui et al., 2015] où le remplacement des mots d'argot, la mise en minuscule et la correction des mots allongés n'ont pas été pertinents, au contraire.

La technique dite "sac de mots" est régulièrement citée. Il s'agit de considérer comme dimensions de l'espace de représentation chacun des termes du corpus. Elle est simple et efficace. [Vincent et Winterstein, 2013] s'en servent pour leur système de classification générique d'opinion, et y ajoutent les traits suivants : présence ou non de négation et catégories grammaticales. Dans [Huang, Murphey, et Ge, 2013], un travail sur le lexique a été effectué afin d'obtenir les termes les plus pertinents au regard des objectifs visés.

b Enrichissement

[Doucy et Massoussi, 2012] abordent l'analyse sémantique des verbatims clients. Des critères et des modèles sont cités pour établir des inférences dans ce type de textes, qui sont présentées comme centrales pour en analyser le contenu et réaliser des traitements automatiques. La catégorisation n'est à l'heure actuelle qu'une perspective de travail des auteurs qui envisagent l'élaboration de grammaires et d'algorithmes pour interpréter des messages. Toutefois, bien qu'ils n'apportent pas de réponse concrète à ce sujet, il est intéressant de noter que l'accent est mis sur la nécessité d'une connaissance et d'une modélisation détaillées du domaine. On peut penser notamment à l'intérêt d'exploiter des ontologies spécialisées. C'est ce que propose par exemple [Maciej, 2008] pour la classification de news sans données d'apprentissage.

L'article de [Chen, Jin, et Shen, 2011] propose une méthode d'enrichissement des documents à partir de sources externes (autres textes portant sur le même sujet, taxonomies) et des lexiques sont encore une fois utilisés dans [Abdaoui et al., 2015] pour ajouter de l'information sur la valeur sémantique des données, qui concernent ici les émotions et les sentiments. Dans [Vernier et al., 2009] l'approche symbolique présentée ainsi que l'enrichissement sur le plan linguistique à l'aide d'un lexique et d'un corpus annoté spécialisés semblent apporter de bons résultats.

D'autres publications vont plus loin encore comme [Eensoo et Valette, 2014] en mettant au premier plan l'apport d'une analyse sémantique textométrique. L'objectif est de représenter les textes par des descripteurs d'ordre sémantique. Enfin, [Janik Kochut, 2007] propose un système de classification basée sur l'exploitation d'ontologies, et qui se passe de données d'entraînement.

c Choix et mise en œuvre d'un algorithme

[Poirier, Fessant, et Tellier, 2010] rapporte des tests de comparaison de différents algorithmes. Le plus performant est un SVM (machine à vecteurs de support) mais il faut prendre en compte le fait que le corpus d'entraînement contient un nombre important de données (170000 commentaires web). C'est également le cas pour la tâche de classification axiologique de tweets de [Abdaoui et al., 2015]. Enfin, au cours de leur expérimentation d'apprentissage automatique avec peu de données d'entrainement, [Salperwyck et Lemaire, 2011] observent que les meilleurs résultats sont obtenus avec des classifieurs naïfs bayésiens.

De façon générale, on voit bien que la qualité des résultats obtenus n'est pas fortement dépendante de l'algorithme, chacun de ceux présentés ayant fait ses preuves, mais plutôt des choix réalisés en amont en terme de représentation des textes.

L'état des connaissances dans le domaine ayant été décrit, nous nous intéressons maintenant aux ressources à notre disposition au sein des structures Holmes et Eloquant pour répondre à notre problématique.

C Analyse de l'existant

L'étape préliminaire pour le développement de notre système de catégorisation générique pour la relation client est de dresser un bilan de l'existant. Pour cela, nous nous intéressons aux méthodes et techniques disponibles, ainsi qu'à un système déjà développé par Holmes pour un contexte spécifique (opérateur de téléphonie et internet). L'analyse de verbatims clients dans ce système aboutit à la catégorisation d'une partie seulement des verbatims. Un nombre important n'est pas classifié faute de correspondre à des règles symboliques ou aux caractéristiques définies par la méthode statistique. Un travail d'analyse de ce système a donc été mené dans le but de dégager des pistes pour le développement de notre futur classifieur générique. En effet, nous nous baserons sur une architecture similaire en faisant appel à des traitements déjà à disposition. Il convient donc de bien en connaître les fonctionnalités et limites.

1. Architecture du système

Le système de classification dont nous disposons est structuré de façon modulaire (il s'agit d'un ensemble de packages Java). Une chaîne de traitement ("pipeline") permet de les mettre en œuvre dans un ordre adéquat et chacun de ces modules peut être paramétré selon les besoins et les données à traiter.

Un certain nombre de traitements linguistiques est déjà à disposition. Il s'agit d'outils externes intégrés à la plate-forme modulaire. L'analyseur syntaxique Talismane[1] développé en Java par Assaf Urieli au sein du laboratoire CLLE-ERSS (Cognition, Langue, Langages, Ergonomie - Équipe de Recherche en Syntaxe et Sémantique) permet de réaliser la segmentation, la tokenisation et l'étiquetage morpho-syntaxique. L'analyse

[1] Urieli, A. (2013). Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit. Université de Toulouse II le Mirail

syntaxique en dépendance est basée sur le corpus French Treebank[2] du Laboratoire de Linguistique Formelle du CNRS. Plusieurs traitements, notamment l'algorithme d'apprentissage automatique pour la classification sont diffusés par le Natural Language Processing Group de l'université de Stanford[3]. Des développements en interne permettent de faire interagir ces modules et adaptant les formats de sorties générées sous forme d'objets Java. Ceci permet d'assurer une interopérabilité des données à chaque étape de la pipeline.

D'autre part, des enrichissements et corrections sont effectués pour optimiser les résultats. Les gazetteers enrichissent les tokens sur le plan sémantique en ajoutant une annotation (par exemple "NEG" est associé au vocable "décevant" pour indiquer une polarité négative). Des règles linguistiques, les tokensregex, permettent d'agir sur des patterns de tokens en ajoutant ou modifiant les annotations ou en participant à la classification (ce point central pour notre travail sera développé dans la deuxième partie). Enfin, les semrules qui portent sur les relations syntaxiques, peuvent être développées pour enrichir sémantiquement l'arbre de dépendance d'une phrase.

Le principe de l'architecture du système ayant été décrit, il s'agit maintenant de relever les points qui font défaut dans le classifieur dont nous disposons.

2. Liste des catégories

La liste des catégories est le premier levier sur lequel il semble judicieux de se pencher. En effet, c'est la base sur laquelle repose le processus de classification. On relève différents éléments susceptibles d'engendrer une catégorisation peu satisfaisante :

- des catégories qui se recoupent (par exemple : "respect des engagements" et "rappel/RDV non honoré") ;

- une disparité de couverture, certaines catégories étant plus vastes que d'autres (par exemple : "accueil-amabilité" est beaucoup plus vaste que "éligibilité aux offres")

- certaines situations non représentées (par exemple : dans l'axe "tonalité du client" trois catégories seulement sont définies et ne couvrent que très partiellement l'axe : "effort client", "risque de perte du client ou de résiliation", "risque juridique") ;

---

[2]Abeillé, A., Clément, L., & Toussenel, F. (2003). Building a Treebank for French. In A. Abeillé (Éd.), *Treebanks* (p. 165-187). Springer Netherlands.
[3] http://nlp.stanford.edu/software/classifier.html

- différents types de catégories qui cohabitent (par exemple : "effort client" relève du ressenti exprimé par le client sur ses démarches. Il s'agit donc de l'expression de la subjectivité, tandis que "renouvellement du matériel" est un point très factuel).

Au-delà de la question de l'automatisation de la tâche, cela pose un problème conceptuel car il est difficile, même pour un opérateur humain d'effectuer une catégorisation systématique sur la base de cette grille. Une attention particulière sera donc accordée à la définition de la liste des catégories, qui doivent être exhaustives au regard du domaine, clairement définies et délimitées.

3. Représentation des documents

Un autre élément important pour la qualité des résultats est la représentation des documents, comme nous l'avons vu dans l'état de l'art. Dans le système qui nous intéresse ici, elle semble très lacunaire car le domaine est mal modélisé. En effet, on ne dispose pas d'un lexique de la relation client pour faire une annotation sémantique lexicale. Ce type de ressource permettrait d'effectuer une représentation des messages avec un certain niveau d'abstraction, bénéfique pour le développement de règles génériques, mais également pour l'analyse statistique. Par exemple, il serait tout à fait pertinent de représenter sous une même annotation des termes comme "société", "compagnie" et "entreprise" car ils renvoient au même concept. Ainsi, la représentation de ce concept via l'annotation lexicale permet de l'exploiter en tant qu'élément unique plutôt qu'en faisant référence à trois termes distincts. On gagne ainsi en généricité.

4. Méthode symbolique

La méthode symbolique du système est constituée de règles établies par un expert linguiste dans le but détecter des patterns spécifiques à chaque catégorie et d'orienter ainsi le classement des verbatims. Or, on constate que les règles actuelles sont très spécifiques car elles se basent sur une représentation de surface des textes : les patterns sont constitués de vocables ou de lemmes mais n'exploitent pas de représentation plus générique, comme les concepts évoqués précédemment, qui regrouperaient plusieurs termes. Les règles actuelles sont une réponse au cas par cas aux problèmes de catégorisation. Un des objectifs de ce travail sera donc de systématiser et de généraliser la méthode symbolique. Cela permettra entre autre de préparer une adaptation aisée à de nouveaux domaines.

5. Evaluation du système

En terme de résultats, on ne dispose pas actuellement de mesures d'évaluation au sein de ce système. Il conviendra donc de mettre en place une méthodologie d'évaluation pour mettre en évidence les apports des développements et des différents volets du système (notamment l'enrichissement sémantique), mais également pour comparer nos résultats à ceux d'autres travaux.

Partie 1

-

Constitution des données

Les données représentent la "matière" de notre système. La fiabilité et la performance de celui-ci dépendent donc en grande partie de leur qualité. La phase de constitution des données est cruciale et une grande attention doit être apportée pour constituer des données structurées et exploitables. Cette première partie est consacrée à la présentation des trois types de données requises pour notre travail : les catégories, les verbatims et les entrées de la taxonomie.

A Définition des catégories

1. Méthodologie

La base du processus de classification est constituée par la grille des catégories. Il convient donc de la définir avec soin afin d'assurer une catégorisation pertinente par la suite. Il s'agit d'un travail conceptuel qu'il ne faut pas négliger car il impacte fortement l'intérêt des résultats, comme nous l'avons constaté dans l'introduction.

Nous proposons la définition suivante pour décrire la notion de catégorie : caractéristique selon laquelle on peut classer un document quand le contenu de celui-ci s'y rapporte. Cette caractéristique peut prendre différentes formes : polarité (négative, positive ou neutre), tonalité (factuelle, menaçante, ...) ou thématique. C'est cette dernière qui nous intéresse car l'objectif poursuivi est une catégorisation générique à la relation client, en fonction des sujets abordés dans les verbatims. Pour un corpus donné, la question à laquelle on souhaite répondre est : de quoi parlent les clients ? Étant donné ce besoin, il faut identifier les éléments qui composent le domaine (ici la relation client) pour obtenir une représentation exhaustive des sujets, et ainsi classer un maximum de verbatims au travers de cette grille.

[Bachimont, 2000] propose une méthodologie de constitution d'une ontologie qui nous semble intéressante pour servir de base à la définition des catégories. En effet, celles-ci peuvent être considérées comme les branches principales d'une ontologie. Nous verrons également dans une prochaine section l'apport de cet article dans notre travail de constitution d'une taxonomie. Ainsi on voit déjà se profiler des liens entre la liste de nos catégories et le lexique structuré en taxonomie.

[Bachimont, 2000] aborde la problématique de la représentation formelle des connaissances et propose une méthodologie pour la construction d'une "ontologie régionale". Sous ce terme, l'auteur désigne une ontologie de domaine (par opposition à une

ontologie universelle), ce qui correspond à notre démarche. De plus, il utilise la notion de "primitives" qui nous intéresse ici car elle va nous permettre d'identifier les éléments fondamentaux du domaine, qui sont en fait les nœuds supérieurs de la représentation hiérarchique de l'ontologie régionale. Enfin, [Bachimont, 2000] insiste sur le fait qu'une ontologie, telle qu'il la propose, dépend du domaine et répond à une tâche, comme c'est notre cas ici. Sa création doit donc impérativement impliquer les acteurs du domaine (on parle d'"approche experte") mais également des linguistes pour l'ingénierie des connaissances. "Une ontologie doit être un consensus des acteurs du domaine." Cette condition a pour but de minimiser au maximum la subjectivité de la représentation des connaissances. Sur la base de cette méthodologie, nous cherchons donc à dresser la liste de nos catégories, avec à l'esprit la nécessité et l'intérêt de représenter le domaine dans une structure hiérarchisée qui nous sera utile lors du développement de la taxonomie.

Nous déterminons ainsi les critères à remplir pour notre travail :

- catégories exhaustives : elles doivent couvrir tout le domaine de la relation client.

- catégories génériques : elles doivent être transversales quel que soit le secteur d'activité.

- catégories objectives : elles doivent permettre, dans la mesure du possible, d'être interprétées de la même manière par les différents acteurs qui s'y référeront. Cela implique donc de choisir des intitulés et d'y joindre une définition et des exemples.

- nombre de catégories exploitable : nous visons une vingtaine de catégories pour représenter le domaine. Il semble en effet difficile au-delà de ce nombre d'exploiter les résultats d'une classification et d'en faire ressortir des tendances. D'autre part, des catégories plus nombreuses seraient nécessairement plus précises et poseraient des difficultés pour le classement de verbatims peu détaillés.

2. Primitives de la relation client : V.1 : à partir des sources existantes

Les différentes listes de catégories existantes au sein des systèmes Holmes servent de point de départ. Un travail de regroupement est effectué pour rassembler celles qui sont proches et/ou se recoupent, et identifier celles qui sont tout à fait distinctes, selon le principe différentiel de [Bachimont, 2000]. Des suppressions (des catégories spécifiques à un secteur d'activité) des fusions, des divisions et des ajouts aboutissent à une première version de la liste des catégories. Celle-ci est affinée à travers un test sur corpus : une

trentaine de verbatims est classée dans cette grille afin de mettre à l'épreuve et d'adapter notre liste.

Axe 1 : Relation client - société

- accueil -> Le client se sent-il bien reçu ? (amabilité, écoute, empathie, confort)
- disponibilité de la société-> Le client arrive-t-il à échanger facilement avec la société ? (attente, ponctualité, proximité, accessibilité, stabilité des interlocuteurs, compréhension...)
- fidélité du client-> Le client va-t-il rester ? (ancienneté, engagement, départ, résiliation, rétractation)
- engagement de la société -> La société tient-elle ses engagements ?
- conseil -> La société est-elle force de conseil ? (support, soutien, aide)
- compétence -> La société et ses agents sont-ils compétents ? (aptitude, efficacité, erreur)
- qualité -> Quelle est la qualité des produits/services de la société ? (simplicité, clarté, efficacité, ergonomie, à jour, éthique, transparence, sécurité, rapidité,...)

Axe 2 : objets et entités du domaine : la société et son offre

- offre (produits, biens, matériel, services)
- tarifs (prix, promotions, geste commercial)
- achat (paiement, facturation, remboursement, commande, livraison, annulation, arnaque, vente forcée)
- contrat
- données et informations
- intervenants et services dans la société
- image de la société - recommandation
- concurrence

Axe 3 : tonalité

- satisfaction
- attente - demande - exigence
- alerte - menace – risque
- RAS

3. Primitives de la relation client : V.2 : approche métier

Suivant les recommandations de [Bachimont, 2000], deux experts de la relation client (directeur du département Solutions Clients et responsable marketing de solutions en relation client) sont consultés afin d'apporter une vision "métier" pour améliorer la grille au regard du domaine. L'objectif de ce travail en groupe est de faire émerger un consensus sur la définition et la délimitation des catégories.

Dans l'axe 1 de la V.1, les catégories semblent trop abstraites, et les sous-thématiques trop nombreuses pour faire office de catégories. Nous cherchons un niveau de précision intermédiaire. Nous privilégions alors une approche de type "parcours client" chronologique (entrée en contact -> prestation -> pérennité de la relation) pour envisager des regroupements et expliciter certains intitulés.

Dans l'axe 2, la seule modification apportée est la suppression de "intervenants et services dans la société" qui semble peu pertinente car elle est abordée à travers les catégories de l'axe 1.

Le troisième axe n'est pas modifié. Nous notons qu'il n'est pas du même ordre que les précédents et nous nous interrogeons sur le fait qu'il relève des "primitives" de la relation client.

Conformément à notre cycle de travail, un nouveau test sur corpus permet de valider cette nouvelle version.

Axe 1 : Relation client – société

- Accueil
- Amabilité (empathie)
- Confort
- Attente – ponctualité

- Accessibilité (proximité, ergonomie)
- Attention portée au client (écoute)
- Stabilité des interlocuteurs
- Support (soutien, aide)
- Proactivité
- Compréhension – intelligibilité
- Compétence (efficacité, rapidité)
- Fidélité du client (ancienneté, engagement)
- Départ (résiliation, rétractation)
- Engagement de la société

Axe 2 : objets et entités du domaine : la société et son offre

- Offre (produits, biens, matériel, services)
- Tarifs (prix, promotions, geste commercial)
- Achat (paiement, facturation, remboursement, commande, livraison, annulation, arnaque, vente forcée)
- Contrat
- Données et informations
- Image de la société – recommandation, confiance, éthique
- Concurrence

Axe 3 : tonalité

- Satisfaction
- Attente - demande – exigence
- Alerte - menace – risque
- RAS

4. Primitives de la relation client : V.3 : approche linguiste

Il s'agit désormais d'envisager la catégorisation du point de vue de la modélisation des connaissances dans un objectif de traitement linguistique. La grille est ainsi analysée par l'équipe d'experts de l'analyse sémantique de Holmes qui a une bonne connaissance du système de catégorisation et plusieurs expériences d'application dans le domaine de la relation client sur des cas spécifiques.

La modification majeure de la grille est la suppression de l'axe 3, pour des raisons pragmatiques. S'agissant plutôt d'un axe sur lequel on se place, il ne relève pas d'une catégorisation thématique. D'autre part, la catégorie "satisfaction" renvoie à l'analyse de sentiment, système à disposition qu'il sera alors possible de croiser avec celui en cours de développement. La catégorie "alerte" est également déjà traitée par la détection de la colère et des risques de résiliation et juridique. Enfin la catégorie "attentes" implique un type d'analyse différent qu'il est à l'heure actuelle difficile de mettre en place de façon satisfaisante.

Sur le plan de la terminologie, nous remplaçons le terme "axe" qui n'est pas adapté ici car il s'agit de thèmes et non effectivement d'"axe" qui indique une polarité ou une quantification. Les catégories sont alors regroupées en "macro-catégories". Il est à noter que ces groupes ne constituent pas des classes de notre système, mais ont simplement une valeur indicative.

Enfin, il est indispensable d'ajouter une catégorie "hors catégorie" pour les verbatims ne portant pas d'information se rapportant à la relation client (par exemple "Bonjour", "rien à signaler", ou encore "pas de commentaire").

Macro-catégorie 1 : Relation client – société

- Accueil
- Attitude émotionnelle de l'agent (amabilité, empathie, écoute)
- Confort
- Attente – ponctualité
- Accessibilité - joignabilité
- Stabilité des interlocuteurs
- Assistance (support, soutien)

- Proactivité
- Compréhension – intelligibilité
- Compétence - efficacité
- Fidélité du client (ancienneté, engagement)
- Départ (résiliation, rétractation)
- Engagement de la société

Macro-catégorie 2 : objets et entités du domaine : la société et son offre

- Offre (produits, biens, matériel, services ; = ce que le client achète)
- Tarifs (prix, promotions, geste commercial)
- Achat (paiement, facturation, remboursement, commande, livraison, annulation, arnaque, vente forcée)
- Contrat
- Données et informations (dont espace client)
- Image de la société (recommandation, confiance, éthique)
- Concurrence
- Hors catégorie

5. Primitives de la relation client : V.4 : intitulés et définitions

La dernière phase pour établir la grille des catégories consiste à choisir les intitulés et à préciser les définitions et périmètres de chaque catégorie. Une attention particulière est portée à la tonalité des intitulés retenus : ceux-ci ne doivent pas sembler trop "polarisés" (par exemple "amabilité" induit une certaine positivité trompeuse).

- **1#accueil** : Première étape du contact entre le client et la société ou ses interlocuteurs.
  "ce qui ma plus c'est l'accueil pas beaucoup de chose mon déplu." "Bon accueil."
- **2#attitude** : Attitude émotionnelle des interlocuteurs de la société. Concerne l'amabilité, la politesse, l'empathie, l'attention, l'écoute.

"vendeuse sympathique et à l'écoute en plus d'être souriante!" "Sur le début conseiller pas très compréhensif et ton ironique ! !!"

- **3#confort** : Bien-être du client procuré par le cadre et les commodités matérielles mis en place par la société.
  "Accueil par le vendeur excellent et très sympathique par contre un peu d'attente dans une boutique un peu petite."

  "L'espace magasin très bien, le personnel à l'écoute et de très bons conseils,visite très agréable."

- **4#attente** : Temps qui s'écoule jusqu'à une action ou un évènement que doit mener la société ou ses interlocuteurs, et attendu par le client. Concerne donc également la ponctualité.

  "l attente un peu longue....."

  "Peu d'attente."

- **5#accessibilité** : Facilité pour le client à entrer en contact avec la société ou ses interlocuteurs. Concerne la disponibilité, la proximité, les horaires.

  "Service client toujour disponible mais mon problème de service blackberry n'etant pas resolu je ne peut pas me permettre de sur estimer la prestation."

  "A qui dois-je adresser mon courrier ?"

- **6#stabilité** : Stabilité ou changement des interlocuteurs de la société auprès du client.

  "Trop d'interlocuteur pour avoir une réponse à notre demande et pendant ce temps nous payons la communication !"

  "J'ai été balladé de services en services pour revenir au final au premier service."

- **7#assistance** : Actions menées par la société et ses interlocuteurs afin de répondre aux attentes et besoins du client. Concerne le support, le soutien, l'aide, le dépannage, la résolution de problème.

  "Je suis très mécontent des résultats de dépannage de votre service."

  "De très bon conseils merci."

- **8#proactivité** : Attitude d'anticipation de problèmes et d'engagement d'évolutions positives.

  "ils ont répondu très vite à ma question e me donnant une nouvelle carte sim qui aille dans le téléphone que l'on m'a donné : parfait! mais personne ne m'a suggéré de sauvegarder mes numéros avant... la carte est donc désactivée et j'ai tout perdu!"

- **9#compréhension** : Capacité des interlocuteurs de la société à se faire comprendre par le client (intelligibilité), ainsi que leur capacité à comprendre ce qu'exprime le client.

  "Et mal à le comprendre avec son très fort accent !!!"

  "Ne comprend aucun mot de ce qu'on lui dit , n'a pas su répondre à mes questions et publicité mensongère !"

- **10#compétence** : Capacité de la société et de ses interlocuteurs à remplir leurs tâches et missions. Concerne également l'efficacité, la rapidité, les bons conseils, le professionnalisme, la qualité de service.

  "La réponse obtenue était très fantaisiste et manifestait un manque sérieux."

  "Bon conseil et personnel agréable et compétent."

- **11#fidélité** : Attitude du client qui, sur la durée, préfère l'offre ou le(s) produit(s) de la société à celle d'un concurrent. Concerne également l'ancienneté.

  "..je suis plus que déçu après plus de 15 ans5chez vous."

  "Etant donné ma fidélité et mon engagement multiple auprès de votre enseigne (j'ai aussi un abonnement box) je voulais savoir si je pouvais bénéficier d'un "effort commercial"."

- **12#départ** : Fin de la relation entre le client et la société. Concerne également la rétractation et la résiliation.

  "Le problème que j'ai avec le réseau [nom de la société : anonymisé] et 4G c pas réglé je suis pas satisfaite je pense résilier mes contrat."

  "Vous perdez 3 clients à la fin 2016!"

- **13#engagement** : Le fait pour la société ou ses interlocuteurs de s'engager à réaliser une ou des actions. Concerne également les promesses.

  "Donc beaucoup de promesse par téléphone au moment de la démarche commercial de vente et plus personne apres ."

  "Ils devaient me rappeler et ne l'ont jamais fait."

- **14#offre** : Ensemble des produits et services que vend la société au client. Concerne aussi la qualité de cette offre.

  "Réseau de mauvaise qualité."

  "Je suis venue m'informer sur les offres de I phone en vue de changer mon mobile."

- **15#tarifs** : Prix à payer par le client. Concerne également les promotions et gestes commerciaux.

  "Coût élevé du forfait."

  "A 20euro alors que je paier que 14.99 euro > ???"

- **16#achat** : Processus au cours duquel le client achète un produit ou un service à la société. Concerne également le paiement, la facturation, le remboursement, la commande, la livraison, l'annulation, l'arnaque, la vente forcée.

  "Je suis content, pasque je demandé ma fature a la maison c'est fait."

  "Mais je suis repartie sans nouveau portable ! "

- **17#contrat** : Lien contractuel entre le client et la société, qui fixe les obligations, les droits et les modalités entre les deux parties.

  "J ai 3 abonnements avec vous."

  "Bonjour, je n'arrive pas à accéder sur le site [nom de la société : anonymisé] à l'ouverture du contrat de renouvellement de mobile, signé électroniquement dans votre magasin [nom d'un partenaire de la société : anonymisé] espace [point de vente de la société : anonymisé], le 29/05/2014. "

- **18#infos** : Données et informations relatives au client, mais également transmises au client. Concerne également l'espace client, les enquêtes.

  "j'ai était bien renseigné, rien ne m'a déplu."

  "Vous pouvez consulter mon dossier pour avoir les détails."

- **19#image** : Image de la société et de ses interlocuteurs auprès du client ou du public. Concerne également la recommandation, la confiance, l'éthique.

  "Rien a branler du client ."

  "Une honte !!! "

- **20#concurrence** : Autres sociétés ou offres similaires, sur le même marché.

  "Dommage que n'y ai pas de cadeau pour un tel achat alors que chez d'autre ça existe."

  "Je souhaite résilier mon abandonnement pour un abonnement sans engagement chez un autre opérateur."

- **21#hors catégorie** : Verbatims ne portant sur aucun élément de la relation client.

  "Tout à était parfait."

  "Flemme."

B Élaboration des corpus de verbatims

Notre système a pour objectif de traiter des verbatims. Au regard de notre objectif, il est nécessaire de disposer de corpus pour plusieurs besoins (décrits précisément dans les sections suivantes) : l'extraction de la terminologie pour la taxonomie, l'entraînement de l'algorithme de la méthode statistique, l'écriture des règles de la méthode symbolique et l'évaluation du système.

1. Sources

Nos corpus proviennent de quatre sources, ayant comme point commun d'appartenir au domaine de la relation client, mais concernant chacune un secteur d'activité différent, afin d’adopter une approche générique à la relation client : banque, mécanique automobile,

opérateur de téléphonie et internet, assurance. Les sources ont les caractéristiques suivantes:

"BANQUE"

Ces verbatims sont issus d'une enquête auprès des clients d'une banque, et sont répartis sur l'année 2015. Il s'agit de réponses à deux questions, la première portant sur les raisons d'une note attribuée à une autre étape de l'enquête, la seconde sur l'amélioration de service. Le canal est ici un formulaire web.

"MECA_AUTO"

Les textes de cette source ont été produits sur les années 2014 et 2015 et concerne également une question portant sur l'amélioration de service dans un formulaire d'enquête web.

"ASSURANCE"

Ici les verbatims sont des avis clients portant sur les assurances auto, habitation et complémentaire santé. Extraits d'un forum de discussion, ils datent de 2008 à 2015.

"TELECOM"

Cette source présente des réponses ouvertes transmises par SMS, suite à une appréciation chiffrée du service lors d'une enquête, en mars 2016.

Notre objectif étant la catégorisation de verbatims, générique à la relation client, il est important de s'appuyer sur des données issues de plusieurs secteurs d'activité différents, tout en ayant en commun les caractéristiques transversales de la relation client : verbatims, récents, portant sur la société et son offre.

2. Format des données

Afin de pouvoir exploiter ces corpus, nous utilisons le format de verbatim défini pour la plate-forme Java (décrite en l'introduction) : les verbatims et leurs métadonnées (par exemple la date, l'origine, etc.) sont donc stockés dans une structure XML qui a une correspondance sous forme d'objet Java. On parle de sérialisation (passage de d'objet Java au XML) et désérialisation (passage inverse) que l'on effectue avec JAXB (Java Architecture for XML Binding). C'est par ces processus qu'il est possible d'extraire et de manipuler les données qui nous intéressent. La structure de chaque document (un verbatim

et ses métadonnées) est la suivante (nous ne présentons ici que les informations que nous exploitons) :

- Document
    - Sentence
        - sentNdx : identifiant de la phrase au sein du document
        - sentenceText : texte de la phrase
    - Concept
        - sentNdx : identifiant de la phrase sur laquelle porte le concept
        - concept : intitulé du concept
    - goldCategories : catégories attribuées manuellement lors de l'étape d'annotation
        - sentNdx : identifiant de la phrase sur laquelle porte la catégorie
        - catName : : intitulé de la catégorie
        - catId : identifiant de la catégorie
    - sysCategories : catégories attribuées automatiquement par le classifieur
        - sentNdx : identifiant de la phrase sur laquelle porte la catégorie
        - catName : intitulé de la catégorie
        - catId : identifiant de la catégorie
        - score : probabilité d'appartenance de la phrase à la catégorie
        - algorithm : méthode de classification (statistique ou hybride)
    - documentText : texte du verbatim
    - corpusId : identifiant du corpus d'origine
    - documentId : identifiant du document au sein du corpus
    - sentenceCount : nombre de phrases dans le document
    - documentDate : date du document

Rappelons ici que notre objectif est de travailler au niveau des phrases, et non des documents.

3. Découpage en sous-corpus

Comme nous l'avons vu, la constitution des corpus doit répondre à différents besoins et étapes de notre travail. Au sein de chaque source, les verbatims ont donc été répartis aléatoirement dans quatre sous-corpus :

"TRAIN"

TRAIN contient des verbatims annotés manuellement pour l'entraînement de l'algorithme d'apprentissage automatique dans le cadre de la méthode statistique.

"TEST"

TEST rassemble des verbatims annotés manuellement pour l'évaluation du système.

"DEVEL"

DEVEL comporte des verbatims qui serviront de base pour l'écriture des règles linguistiques dans le cadre de la méthode symbolique. Ces données ne sont pas annotées.

Les étapes de développement des méthodes et d'évaluation seront décrites en deuxième et troisième parties, mais il est important de noter ici la nécessité de constituer des sous-corpus distincts afin d'éviter tout phénomène de "corpus over-fitting". En effet, si deux ou trois de ces étapes s'appuient sur les mêmes données, les performances du système seront augmentées de façon artificielle. Par exemple, si l'on évalue le système sur les données avec lesquelles l'algorithme de la méthode statistique est entraînée, les résultats seront nécessairement corrects.

"RESTE"

RESTE contient tous les autres verbatims qui n'entrent pas dans ces sous-corpus mais qui seront exploités pour l'extraction de la terminologie.

En terme de volumétrie, les quantités suivantes (en nombre de phrases) ont été retenues pour chaque corpus : 50 pour le TEST, 450 pour le TRAIN, 500 pour le DEVEL. La principale contrainte étant le temps à consacrer à l'annotation manuelle, il nous semble que 2000 phrases (réparties entre 50 pour le TEST et 450 pour le TRAIN, multipliés par quatre sources) est le maximum. Pour effectuer cette répartition, une classe Java (XmlDocumentManaging.java) a été spécifiquement développée (cf. Annexe 1 : Développements informatiques réalisés).

Le tableau suivant présente la répartition des phrases et verbatims par source et par sous-corpus.

	BANQUE		MECA_AUTO		ASSURANCE		TELECOM	
	Nb verbatims	Nb phrases	Nb verbatims	Nb phrases	Nb verbatims	Nb phrases	Nb verbatims	Nb phrases
TEST	31	50	44	50	21	61	36	50
TRAIN	348	450	379	450	126	450	285	450
DEVEL	405	500	406	500	145	500	342	500
RESTE	11478	14796	14799	18067	21604	72083	35253	48864
Total	12262	15796	15628	19067	21896	73094	35916	49864
Nb phrases /verbatim	1,29		1,22		3,34		1,39	
Nb tokens /phrase	14,47		13,94		21,89		12,73	
Nb caractères /phrase	119,59		108,19		388,93		102,77	
Nb caractères /token	8,27		7,76		17,77		8,08	

Figure 3: Répartition des phrases, verbatims, caractères et tokens dans les corpus

Il faut noter que les phrases d'un même verbatim n'étant pas séparées lors du découpage en sous-corpus, il peut arriver que les quantités effectives de phrases dépassent nos quantités fixées plus haut, sans que cela ne constitue un biais pour notre système.

Un point intéressant à relever est le nombre de phrases par verbatim en fonction de la source. Alors que pour BANQUE, MECA_AUTO et TELECOM, il se situe entre 1,22 et 1,39, il est nettement plus élevé pour ASSURANCE. Cela s'explique probablement par le type de support (forum web) qui incite moins à la concision que les canaux SMS ou formulaire web. D'autre part, s'agissant d'avis web déposés dans une démarche proactive de la part du client, on peut supposer que le client a "plus" à dire que celui qui est sollicité dans le cadre d'une enquête. Ici encore, comme nous faisons le choix de travailler au niveau des phrases, cela ne devrait pas avoir d'impact direct sur les performances du classifieur. Par contre, il faut reconnaître que plus un verbatim contient de phrases, plus il risque d'y avoir des inférences implicites qui ne peuvent être détectées et interprétées par le système.

On constate que le phénomène est le même pour le ratio caractères par phrase, tokens (mots et ponctuations) par phrase, et caractères par token. Les documents de la source ASSURANCE contiennent plus de mots et les mots sont plus longs que dans les autres sources. Les raisons que nous pouvons avancer sont les mêmes que celle évoquées précédemment pour le nombre de phrases par verbatim. Cependant, cette quantité

supérieure d’information viendra peut-être au contraire améliorer les performances en terme de classification.

4. Annotation

Deux sous-corpus requièrent une annotation manuelle : TEST et TRAIN. Il s'agit d'établir une classification qui servira ensuite à entraîner l'algorithme d'apprentissage automatique (le sous-corpus TRAIN constituant alors un modèle) et à évaluer le classifieur (le sous-corpus TEST permet de comparer les résultats obtenus automatiquement avec la classification réalisée par un expert humain).

Pour ce faire, la méthodologie est la suivante : pour chaque phrase, une ou plusieurs catégories sont attribuées en se basant sur le contenu exprimé. Deux contraintes doivent être respectées. Tout d'abord, dans le cas où il y a plusieurs catégories, il faut identifier celle qui semble la plus importante car c'est elle uniquement qui sera utilisée comme donnée d'entraînement. De plus, il faut autant que possible être exhaustif afin d'attribuer toutes les catégories que couvre la phrase, cela permettra une évaluation précise.

Les tableaux suivants présentent la répartition de l'annotation manuelle, c’est-à-dire le nombre de phrases annotées par catégorie, par sous-corpus et par source. Pour extraire ces données, une classe Java (AnnotationStat.java) a été spécifiquement développée (cf. Annexe 1 : Développements informatiques réalisés).

Dans le premier, on trouve le détail pour chaque ensemble source/sous-corpus, les catégories étant classées dans l'ordre de fréquence décroissant sur l'ensemble des corpus.

	TEST					TRAIN					Total TEST + TRAIN				
#categories	BANQUE	MECA_ AUTO	ASSU RANCE	TELECOM	Total	BANQUE	MECA_ AUTO	ASSU RANCE	TELECOM	Total	BANQUE	MECA_ AUTO	ASSU RANCE	TELECOM	Toutes Sources
21#HORSCATEG	9	10	11	8	38	85	119	72	116	392	94	129	83	124	430
10#COMPETENCE	11	12	11	12	46	98	61	71	79	309	109	73	82	91	355
2#ATTITUDE	5	7	4	6	22	93	33	29	66	221	98	40	33	72	243
15#TARIFS	5	5	18	3	31	23	50	106	11	190	28	55	124	14	221
14#OFFRE	2	5	3	8	18	33	47	35	41	156	35	52	38	49	174
5#ACCESSIBILITE	8	1	4	2	15	73	18	37	18	146	81	19	41	20	161
16#ACHAT	1	4	10	3	18	14	32	36	38	120	15	36	46	41	138
1#ACCUEIL	4	2	1	7	14	50	17	8	38	113	54	19	9	45	127
4#ATTENTE	5	6	1	6	18	10	40	18	28	96	15	46	19	34	114
18#INFOS	2	1	2	4	9	23	28	22	30	103	25	29	24	34	112
19#IMAGE	6	0	4	0	10	17	7	60	10	94	23	7	64	10	104
7#ASSISTANCE	0	0	1	3	4	11	3	16	56	86	11	3	17	59	90
11#FIDELITE	0	1	4	0	5	21	11	31	9	72	21	12	35	9	77
12#DEPART	0	0	6	1	7	6	1	43	17	67	6	1	49	18	74
3#CONFORT	0	9	0	0	9	1	47	0	2	50	1	56	0	2	59
17#CONTRAT	1	0	4	0	5	4	0	34	4	42	5	0	38	4	47
20#CONCURRENCE	0	0	10	2	12	10	2	15	5	32	10	2	25	7	44
6#STABILITE	1	0	2	0	3	15	4	7	10	36	16	4	9	10	39
13#ENGAGEMENT	0	0	1	4	5	2	5	6	14	27	2	5	7	18	32
9#COMPREHENSION	1	1	2	3	7	1	0	5	15	21	2	1	7	18	28
8#PROACTIVITE	2	1	0	1	4	5	1	0	0	6	7	2	0	1	10
Total	63	65	99	73	300	595	526	651	607	2379	658	591	750	680	2679

Figure 4 : Nombre de phrases annotées dans chaque catégorie par source/sous-corpus, triées sur la fréquence totale

La première remarque que l'on peut faire est que la catégorie 21#HORSCATEG est la plus présente. On a donc un nombre important de phrases qui ne relèvent d'aucun des centres d'intérêt de notre catégorisation. En effet, beaucoup de clients donnent des réponses peu précises ou indiquant uniquement sur leur satisfaction sans préciser sur quoi celle-ci porte. Trois catégories semblent centrales : 10#COMPETENCE, 2#ATTITUDE et 15#TARIFS sont les thématiques les plus largement abordées dans notre corpus. A l'inverse, quelques catégories semblent plus marginales. C'est le cas notamment de 8#PROACTIVITE. On voit d'ores-et-déjà se dessiner ici des tendances qui permettent de répondre à notre question initiale, but de notre classifieur : de quoi parlent les clients ?

Dans le deuxième tableau qui reprend une partie des données du précédent, les fréquences absolues et relatives sont calculées sur chaque source (les sous-corpus sont alors regroupés) et classées dans l'ordre de fréquence décroissante pour chacune de ces sources.

BANQUE			MECA_AUTO		
#categories	fréquence	fréquence relative	#categories	fréquence	fréquence relative
10#COMPETENCE	109	0,17	21#HORSCATEG	129	0,22
2#ATTITUDE	98	0,15	10#COMPETENCE	73	0,12
21#HORSCATEG	94	0,14	3#CONFORT	56	0,09
5#ACCESSIBILITE	81	0,12	15#TARIFS	55	0,09
1#ACCUEIL	54	0,08	14#OFFRE	52	0,09
14#OFFRE	35	0,05	4#ATTENTE	46	0,08
15#TARIFS	28	0,04	2#ATTITUDE	40	0,07
18#INFOS	25	0,04	16#ACHAT	36	0,06
19#IMAGE	23	0,03	18#INFOS	29	0,05
11#FIDELITE	21	0,03	5#ACCESSIBILITE	19	0,03
6#STABILITE	16	0,02	1#ACCUEIL	19	0,03
16#ACHAT	15	0,02	11#FIDELITE	12	0,02
4#ATTENTE	15	0,02	19#IMAGE	7	0,01
7#ASSISTANCE	11	0,02	13#ENGAGEMENT	5	0,01
20#CONCURRENCE	10	0,02	6#STABILITE	4	0,01
8#PROACTIVITE	7	0,01	7#ASSISTANCE	3	0,01
12#DEPART	6	0,01	20#CONCURRENCE	2	0,00
17#CONTRAT	5	0,01	8#PROACTIVITE	2	0,00
13#ENGAGEMENT	2	0,00	12#DEPART	1	0,00
9#COMPREHENSION	2	0,00	9#COMPREHENSION	1	0,00
3#CONFORT	1	0,00	17#CONTRAT	0	0,00
Total	658		Total	591	

ASSURANCE			TELECOM		
#categories	fréquence	fréquence relative	#categories	fréquence	fréquence relative
15#TARIFS	124	0,17	21#HORSCATEG	124	0,18
21#HORSCATEG	83	0,11	10#COMPETENCE	91	0,13
10#COMPETENCE	82	0,11	2#ATTITUDE	72	0,11
19#IMAGE	64	0,09	7#ASSISTANCE	59	0,09
12#DEPART	49	0,07	14#OFFRE	49	0,07
16#ACHAT	46	0,06	1#ACCUEIL	45	0,07
5#ACCESSIBILITE	41	0,05	16#ACHAT	41	0,06
14#OFFRE	38	0,05	4#ATTENTE	34	0,05
17#CONTRAT	38	0,05	18#INFOS	34	0,05
11#FIDELITE	35	0,05	5#ACCESSIBILITE	20	0,03
2#ATTITUDE	33	0,04	12#DEPART	18	0,03
20#CONCURRENCE	25	0,03	13#ENGAGEMENT	18	0,03
18#INFOS	24	0,03	9#COMPREHENSION	18	0,03
4#ATTENTE	19	0,03	15#TARIFS	14	0,02
7#ASSISTANCE	17	0,02	19#IMAGE	10	0,01
1#ACCUEIL	9	0,01	6#STABILITE	10	0,01
6#STABILITE	9	0,01	11#FIDELITE	9	0,01
13#ENGAGEMENT	7	0,01	20#CONCURRENCE	7	0,01
9#COMPREHENSION	7	0,01	17#CONTRAT	4	0,01
3#CONFORT	0	0,00	3#CONFORT	2	0,00
8#PROACTIVITE	0	0,00	8#PROACTIVITE	1	0,00
Total	750		Total	680	

Figure 5: Nombre de phrases annotées dans chaque catégorie par source et fréquence relative, triées par ordre décroissant dans chaque source

Un comparatif par source montre des spécificités par secteur d'activité. Alors que 10#COMPETENCE reste en tête dans chaque source, on constate que 2#ATTITUDE passe du haut de la liste sur BANQUE et TELECOM à une position intermédiaire sur MECA_AUTO et ASSURANCE. D'autre part, 9#COMPREHENSION présente une fréquence nettement plus importante sur la source TELECOM que sur les autres, cela s'explique sans doute par le fait que cette société dispose de centres d'appels à l'étranger : la question de la compréhension et de l'intelligibilité est donc récurrente dans le discours des clients. Il sera intéressant de vérifier ses spécificités dans les résultats obtenus à l'issu du développement du classifieur.

Le dernier tableau met en avant le nombre de catégories attribuées par phrase pour chaque ensemble source/sous-corpus. Une distinction est faite pour exclure ou non 21#HORSCATEG qui n'a pas le même rôle que les autres.

		1 catégorie		1 catégorie (hors 21)		2 catégories		3 catégories		4 catégories		5 catégories		
	source	fréquence	fréquence relative	fréquence	fréquence relative	fréquence	fréquence relative	fréquence	fréquence relative	fréquence	fréquence relative	fréquence	fréquence relative	Total (dont 21)
TEST	BANQUE_TEST	38	0,760	29	0,580	11	0,220	1	0,020	0	0,000	0	0,000	50
	MECA_AUTO_TEST	36	0,720	26	0,520	13	0,260	1	0,020	0	0,000	0	0,000	50
	ASSURANCE_TEST	35	0,574	24	0,393	15	0,246	10	0,164	1	0,016	0	0,000	61
	TELECOM_TEST	30	0,600	22	0,440	18	0,360	1	0,020	1	0,020	0	0,000	50
TRAIN	BANQUE_TRAIN	336	0,743	251	0,555	91	0,201	23	0,051	2	0,004	0	0,000	452
	MECA_AUTO_TRAIN	391	0,861	272	0,599	55	0,121	7	0,015	1	0,002	0	0,000	454
	ASSURANCE_TRAIN	309	0,685	237	0,525	96	0,213	36	0,080	8	0,018	2	0,004	451
	TELECOM_TRAIN	313	0,705	197	0,444	104	0,234	22	0,050	5	0,011	0	0,000	444
TEST + TRAIN	BANQUE_total	374	0,745	280	0,558	102	0,203	24	0,048	2	0,004	0	0,000	502
	MECA_AUTO_total	427	0,847	298	0,591	68	0,135	8	0,016	1	0,002	0	0,000	504
	ASSURANCE_total	344	0,672	261	0,510	111	0,217	46	0,090	9	0,018	2	0,004	512
	TELECOM_total	343	0,694	219	0,443	122	0,247	23	0,047	6	0,012	0	0,000	494
	total	1488	0,740	1058	0,526	403	0,200	101	0,050	18	0,009	2	0,001	2012

Figure 6 : Nombre de phrases en fonction du nombre de catégories attribuées et fréquence relative pour chaque source/sous-corpus

La première observation que l'on fait est que la majorité des phrases porte une seule catégorie : 74% des phrases (dont 53% autres que 21#HORSCATEG), tandis que très peu renvoient à plus de trois catégories. La source ASSURANCE confirme notre hypothèse évoquée plus haut : contenant plus d'informations, le nombre de catégories attribuées est plus important que sur les autres sources.

C Création d'une taxonomie

Le troisième type de données à préparer pour notre travail concerne le lexique : il s'agit d'élaborer une taxonomie de la relation client.

1. Objectif de représentation des documents

Comme nous l'avons vu en introduction, la représentation des documents est un point important pour une tâche de catégorisation et notre hypothèse de travail est l'amélioration des performances d'un classifieur par l'enrichissement sémantique. Le principe est donc de réaliser une annotation lexicale des verbatims qui permette une représentation des concepts exprimés, au sein du domaine qui nous intéresse, à savoir la relation client. Ainsi, chaque vocable relatif au domaine est associé à une annotation qui indique sa place dans notre représentation du domaine. Il s'agit donc de passer du lexique aux concepts du domaine, hiérarchisés dans une taxonomie. Cette structure permet de décrire les liens qu'entretiennent entre eux les différents concepts et le lexique qui les représente dans les documents. L'enrichissement sémantique des textes, comme nous l'avons vu dans l'état de l'art, apporte de meilleurs résultats de classification, particulièrement sur des textes courts.

Notre méthodologie pour la création de cette taxonomie s'appuie sur trois publications que nous présentons dans la suite de cette section. Celles-ci traitent de la constitution d'ontologies, qui sont des représentations plus complexes mêlant liens taxonomiques (synonymie, méronymie, hyperonymie, par exemple "X est un homme", "un homme est un être humain" et "Y est un film" ; "un film est une œuvre") et liens sémantiques (de type : "X est acteur de Y"). Nous nous limiterons ici aux premiers. C'est pourquoi nous parlons de "taxonomie", qui constitue la première couche d'une ontologie.

2. Extraction de la terminologie

[Mondary, 2008] aborde la construction d'ontologie sur la base de corpus. L'intérêt de s'appuyer sur des usages attestés est mis en avant. Dans notre cas, cela semble primordial pour adopter une représentation la plus fidèle possible du domaine auquel appartiennent les textes que nous voulons traiter. Les trois étapes que présente [Mondary, 2008] et que nous nous proposons de suivre sont les suivantes :

- identification des termes ;

- regroupement en classes sémantiques ;

- structuration en réseau terminologique.

Nous commençons donc par une analyse terminologique sur l'ensemble des corpus préparés, tels que nous les avons décrits dans la section précédente. Pour cela, nous faisons le choix d'avoir recours à la fonctionnalité "extraction de concepts" intégrée à la plate-forme Holmes utilisant le modèle proposé par [Sclano et Velardi, 2007]. Ce système, qui compare les fréquences d'apparition des termes entre le corpus spécifique et un corpus généraliste, permet de détecter la terminologie propre au domaine qui nous intéresse.

Par "concept", on entend ici les termes et expressions qui sont analysés comme majeurs pour indiquer le contenu sémantique du verbatim. Afin d'éviter une confusion avec la notion de concept évoquée jusqu'ici et par la suite, nous utiliserons la dénomination "terme". Les objets Java décrivant les verbatims contiennent les termes identifiés par le système dans la propriété "concept". Des développements Java (ConceptValues.java, XmlConceptOutput.java, XmlConceptRetrieval.java) ont été réalisés (cf. Annexe 1 : Développements informatiques réalisés) pour obtenir une sortie présentant les termes extraits accompagnés de leur fréquence dans chaque sous-corpus. Nous obtenons ainsi une liste de 144 089 termes.

SOURCE :	MECA_ AUTO	MECA_ AUTO	TELECOM	TELECOM	BANQUE	BANQUE	ASSU RANCE	ASSU RANCE
terme	fréquence	fréquence relative	fréquence	fréquence relative	fréquence	fréquence relative	fréquence	fréquence relative
charlotte	0	0,000000	1	0,000009	0	0,000000	0	0,000000
charmant	0	0,000000	1	0,000009	0	0,000000	0	0,000000
charmant assureur	0	0,000000	0	0,000000	0	0,000000	1	0,000003
charmant conseiller	0	0,000000	2	0,000019	0	0,000000	2	0,000007
charmant conseiller à top	0	0,000000	1	0,000009	0	0,000000	0	0,000000
charmant courrier	0	0,000000	0	0,000000	0	0,000000	1	0,000003
charmant courrier de recouvreme	0	0,000000	0	0,000000	0	0,000000	1	0,000003
charmant dame	1	0,000016	1	0,000009	0	0,000000	0	0,000000
charmant demoiselle	1	0,000016	1	0,000009	0	0,000000	0	0,000000
charmant hôtesse	0	0,000000	1	0,000009	0	0,000000	0	0,000000
charmant jeune fille	0	0,000000	1	0,000009	0	0,000000	0	0,000000
charmant jeune homme	0	0,000000	1	0,000009	0	0,000000	0	0,000000
charmant personne	0	0,000000	1	0,000009	0	0,000000	1	0,000003
charmant vendeur	0	0,000000	2	0,000019	0	0,000000	0	0,000000
charme	0	0,000000	2	0,000019	0	0,000000	0	0,000000
charme de la conseiller	0	0,000000	1	0,000009	0	0,000000	0	0,000000
charment jeune homme	0	0,000000	1	0,000009	0	0,000000	0	0,000000
charmeur	0	0,000000	0	0,000000	0	0,000000	1	0,000003
charpente	0	0,000000	0	0,000000	0	0,000000	2	0,000007
charpente de mur	0	0,000000	0	0,000000	0	0,000000	1	0,000003
charrette	0	0,000000	0	0,000000	0	0,000000	1	0,000003
charrette de 1999	0	0,000000	0	0,000000	0	0,000000	1	0,000003
charte	2	0,000032	0	0,000000	1	0,000019	4	0,000013
charte agréable	0	0,000000	0	0,000000	1	0,000019	0	0,000000
charte client	1	0,000016	0	0,000000	0	0,000000	0	0,000000
charte de bon conduite	0	0,000000	0	0,000000	0	0,000000	1	0,000003
charte qualité	1	0,000016	0	0,000000	0	0,000000	2	0,000007

Figure 7 : Extrait de l'extraction des termes avec leurs fréquence par source

La principale limite de cette fonctionnalité "extraction de concepts" est qu'elle porte uniquement sur des groupes nominaux. Nous considérons qu'elle constitue néanmoins une bonne base et que nous pourrons élargir ultérieurement le processus aux verbes et adjectifs notamment.

Une fois les termes extraits, il est nécessaire d'en faire une sélection. Après avoir éliminé le bruit, il faut en effet distinguer ceux qui relèvent du domaine de la relation client en général de ceux qui concernent les secteurs d'activité de nos quatre sources (par exemple "garage", bien qu'ayant une fréquence élevée, n'est en fait pertinent que pour le secteur de la source MECA_AUTO). Dans un premier temps, nous filtrons les termes présents dans au moins trois sources et ayant une fréquence relative (sur l'ensemble des corpus) supérieure à 0.0005. 405 termes sont retenus par ce filtre. Nous établissons et comparons alors plusieurs tris à partir des fréquences dans le but de choisir le plus adapté afin de sélectionner les entrées qui correspondent à notre besoin : moyenne des fréquences relatives des quatre sources, écart-type et coefficient de variation. Une observation de ces différentes mesures révèle que le tri le plus pertinent semble être l'ordre décroissant de la moyenne des fréquences de chaque source. A ce stade, un tri manuel, faisant appel à l'expertise du linguiste en matière de représentation des connaissances, est requis pour

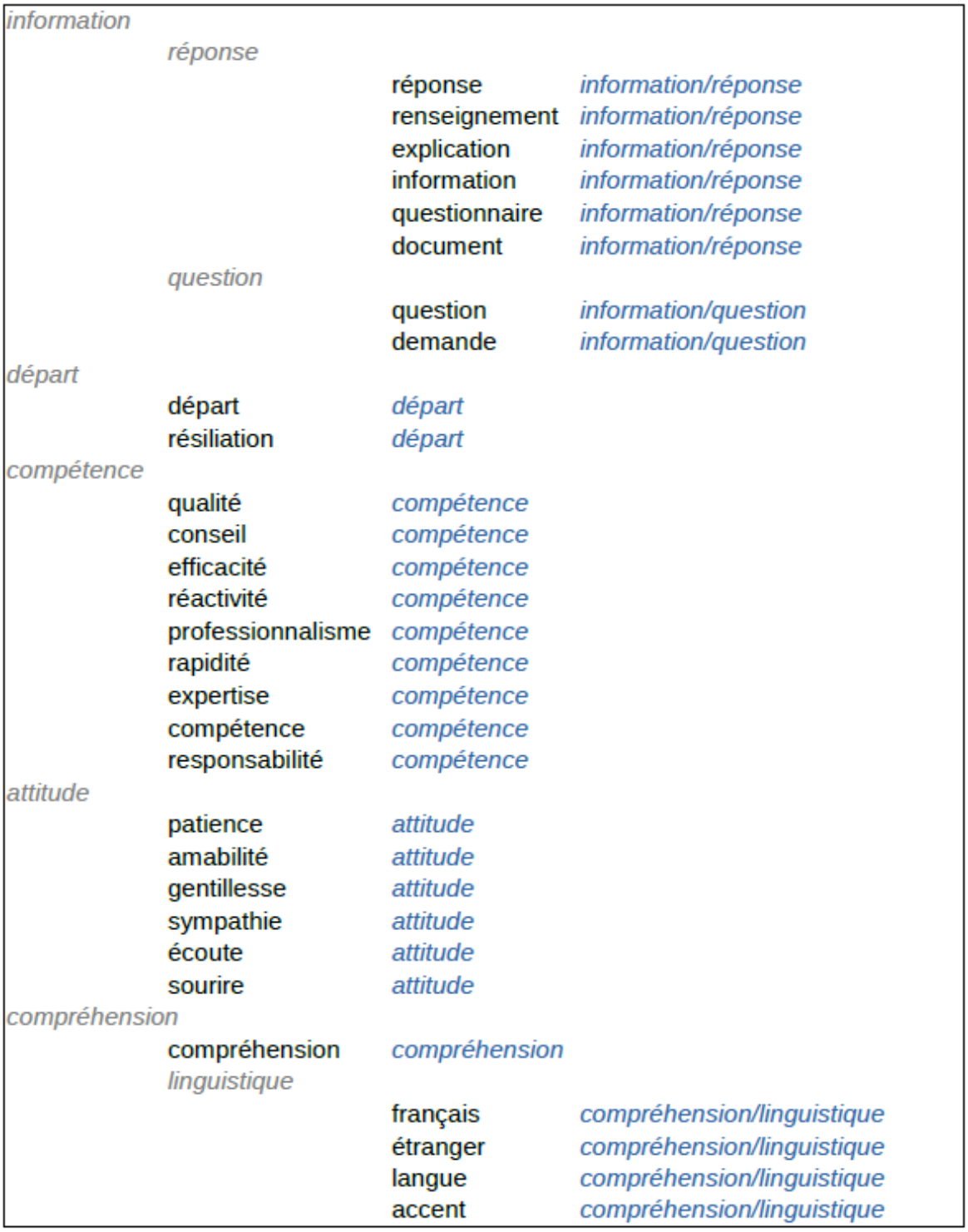
sélectionner les termes qui constitueront la taxonomie de la relation client. Il s'agit d'éliminer ceux qui semblent plus larges que le domaine cible (par exemple "moment") et ceux qui sont trop spécifiques à une de nos sources (par exemple "véhicule"). Nous obtenons ainsi 214 termes.

3. Organisation en structure hiérarchisée

L'étape suivante proposée par [Mondary, 2008] est le regroupement en classes sémantiques. Ici encore il s'agit d'un travail relevant de l'expertise du linguiste : nous procédons à des rapprochements de termes, des suppressions de doublons et des simplifications. Par exemple "conseils", "très bons conseils" et "bons conseils" sont regroupés et aboutissent à "conseil" ; "langue" et "accent" sont rapprochés ; etc.

Ce travail s'appuie également sur [Bachimont, 2000] que nous avons déjà abordé dans la première section. En effet, nous réalisons ce regroupement en classes sémantiques en ayant à l'esprit la liste de nos catégories de classification. Comme nous l'avons vu, ces catégories peuvent constituer le premier niveau (c’est-à-dire les branches principales) de la taxonomie qui a pour objectif de représenter le domaine, à la fois pour la catégorisation, et pour l'enrichissement sémantique, en vue de cette même catégorisation. Ainsi, nous distinguons les termes qui relèvent de la compétence de ceux qui relèvent de l'attitude, bien que ceux-ci, dans une autre perspective, auraient pu tout à fait être rapprochés.

De plus, la structure de notre taxonomie répond aux caractéristiques décrites dans [Bachimont, 2000], à savoir les principes d'identité et de différences entre les concepts. Un système d'héritage de propriétés (les sèmes, c’est-à-dire des traits sémantiques, constituants de la signification globale d'une unité) met en évidence les liens qui existent. Entre une unité mère et sa fille, il y a communauté de tous les sèmes de la mère, et un trait vient préciser le sens de la fille. Entre deux unités sœurs, il y a communauté des sèmes issus de l'unité mère, et un trait supplémentaire différent pour chaque sœur. Nous cherchons ainsi à constituer notre base de connaissances, en organisant une taxonomie de concepts dans laquelle nous plaçons, au niveau des nœuds ou des feuilles, les termes issus de notre analyse et sélection terminologiques. Le recours à une base de connaissance est d'ailleurs selon [Charlet et Bachimont, 2004] un préalable à tout système lorsque l'on veut réaliser des traitements intelligents au regard d'un domaine.



information			
	réponse		
		réponse	*information/réponse*
		renseignement	*information/réponse*
		explication	*information/réponse*
		information	*information/réponse*
		questionnaire	*information/réponse*
		document	*information/réponse*
	question		
		question	*information/question*
		demande	*information/question*
départ			
	départ	*départ*	
	résiliation	*départ*	
compétence			
	qualité	*compétence*	
	conseil	*compétence*	
	efficacité	*compétence*	
	réactivité	*compétence*	
	professionnalisme	*compétence*	
	rapidité	*compétence*	
	expertise	*compétence*	
	compétence	*compétence*	
	responsabilité	*compétence*	
attitude			
	patience	*attitude*	
	amabilité	*attitude*	
	gentillesse	*attitude*	
	sympathie	*attitude*	
	écoute	*attitude*	
	sourire	*attitude*	
compréhension			
	compréhension	*compréhension*	
	linguistique		
		français	*compréhension/linguistique*
		étranger	*compréhension/linguistique*
		langue	*compréhension/linguistique*
		accent	*compréhension/linguistique*

Légende		
concept	terme	*place du terme dans la taxonomie des concepts*

Figure 8 : Extrait de la taxonomie à partir des termes de l'extraction automatique

A ce stade nous disposons d'une structure hiérarchisée de concepts dans laquelle nous avons placés un certain nombre de termes (uniquement des noms) du domaine de la

relation client. Nous étoffons cette base, notamment avec des adjectifs et des verbes, à travers l'exploration du corpus DEVEL (auquel nous avons recours dans le cadre du développement de la méthode symbolique, décrite dans la deuxième partie). En effet, en nous inspirant d'exemples de verbatims non ou mal classés, nous identifions de nouveaux termes pertinents pour notre taxonomie. Il serait bien entendu trop ambitieux, et impossible, de viser l'exhaustivité, mais un développement de cette ressource lexicale au fur et à mesure de nos observations des résultats du classifieur nous semble très bénéfique.

Concrètement, la taxonomie est représentée au sein de notre système dans plusieurs fichiers. Les vocables uniques sont enregistrés dans des "gazetteers". Un gazetteer différent est utilisé pour chaque différente "part-of-speech" (catégorie syntaxique). Ainsi le terme "français" se trouve dans le gazetteer des adjectifs : il n'est enrichi avec son annotation (COMPREHENSION/LINGUISTIQUE) que lorsqu'il est analysé par l'analyseur syntaxique comme adjectif, et non comme nom. Cela permet une première forme de désambiguïsation.

Les expressions polylexicales, figées ou non, n'étant pas traitées par les gazetteers, sont gérées par les tokensregex qui ont pour but de détecter des patterns syntaxiques (le formalisme de ces règles est décrit précisément dans la deuxième partie). De même, les termes ambigus que l'on peut désambiguïser par des éléments de contexte dans la phrase sont gérés de la même façon.

Afin de permettre une exploration plus aisée dans ces fichiers de lexique, deux développements (TaxonomyManaging.java, create_taxonomy.sh) ont été réalisés (cf. Annexe 1 : Développements informatiques réalisés).

Partie 2

-

Développement de la méthode hybride de classification automatique

Le développement d'un classifieur avec enrichissement sémantique, tel que nous l'avons défini dans nos objectifs, se base sur la combinaison de deux méthodes : statistique, qui constitue ici la base du système, et symbolique, qui vient en complément dans un but de correction et d'amélioration de la première. Pour cela, nous procédons par cycle de développement et d'évaluation.

A Représentation des documents

Un travail préliminaire est mené sur la représentation des documents. Il s'agit de mettre en place différents traitements linguistiques sur les données au sein d'une pipeline permettant de décrire chaque document selon des caractéristiques structurées et exploitables. Les traitements retenus pour analyser les documents sont les suivants :

- tokenisation

- analyse morpho-syntaxique

- analyse syntaxique en dépendance

Des modules complémentaires permettent d'affiner ces traitements et de corriger un certain nombre d'erreurs liées à l'ambiguïté ou à une orthographe non standard.

A ces analyses traditionnelles en classification de textes, nous ajoutons l'annotation lexicale sémantique qui fait appel à la taxonomie (décrite dans la première partie). En effet, nous cherchons à savoir quel est l'impact sur les performances d'un classifieur de la représentation des documents sur le plan sémantique, en associant les vocables relevant de la terminologie du domaine à une représentation structurée de celui-ci.

Un exemple de verbatim analysé dans cette pipeline est présenté ci-dessous : "Je n'ai pas attendu et l'hôtesse a été a l'écoute et très professionnelle."

Id	Word	Lemma	POS	SEM-A
1	Je	je	CLS	
2	n'	ne	ADV	
3	ai	avoir	V	
4	pas	pas	ADV	
5	attendu	attendre	VPP	ATTENDRE
6	et	et	CC	
7	l'	le	DET	
8	hôtesse	hôtesse	NC	SOCIETE#ENTITE#INTERLOCUTEUR
9	a	avoir	V	
10	été	être	VPP	
11	à	à	P	
12	l'	le	DET	
13	écoute	écoute	NC	ATTITUDE
14	et	et	CC	
15	très	très	ADV	
16	professionnelle	professionnel	ADJ	
17	.	.	PONCT	

- suj (attendu-5 , Je-1)
- mod (attendu-5 , n'-2)
- aux_pass (attendu-5 , ai-3)
- mod (attendu-5 , pas-4)
- root (root-0 , attendu-5)
- coord (attendu-5 , et-6)
- det (hôtesse-8 , l'-7)
- suj (été-10 , hôtesse-8)
- aux_tps (été-10 , a-9)
- dep_coord (et-6 , été-10)
- mod (été-10 , à-11)
- det (écoute-13 , l'-12)
- obj (à-11 , écoute-13)
- coord (à-11 , et-14)
- mod (professionnelle-16 , très-15)
- dep_coord (et-14 , professionnelle-16)
- ponct (attendu-5 , .-17)

Figure 9 : Analyse d'un verbatim dans la pipeline de représentation d'un document

Chacune des analyses effectuées est ensuite exploitable dans les méthodes statistique et symbolique.

B Méthode statistique

Comme cela a été décrit dans l'état de l'art, la méthode statistique comporte deux phases. Lors de la première, un apprentissage automatique sur des données annotées (ici notre corpus TRAIN) par l'algorithme choisi permet de construire un modèle de classification. Il est donc ensuite possible (deuxième phase) de procéder à la classification de données nouvelles en entrée. L'algorithme utilisé, de type classifieur linéaire multi-classe, est celui développé par le Natural Language Processing Group de l'université de Stanford et est particulièrement adapté à une tâche de classification de texte.

1. Choix des traits

Les deux phases évoquées recourent à la représentation des documents via des traits. Sur le plan fonctionnel, la méthode statistique prend en entrée la sortie de notre pipeline de représentation des documents. Il est bien entendu indispensable que les documents soient représentés de la même façon, c'est-à-dire par les mêmes traits, lors des deux phases. A partir de la représentation des documents décrite à la section précédente, nous choisissons les traits suivants pour la méthode statistique :

- Word : la forme des tokens tels qu'ils apparaissent dans le verbatim
- Lemma : la forme lemmatisée des tokens
- POS (part-of-speech) : la catégorie syntaxique des tokens
- Les dépendances syntaxiques sous forme de triplet : type de relation (premier token, deuxième token)
- Les groupes nominaux (lorsqu'ils sont composés d'au moins deux tokens), issus de l'analyse syntaxique en dépendance.
- L'annotation sémantique des tokens le cas échéant

Le choix de ces traits reprend la sélection présente dans le système de classification déjà en place sur la plateforme Holmes et qui nous semble pertinente au regard de l'état de l'art. Nous y ajoutons l'annotation sémantique basée sur la taxonomie car c'est cet aspect que nous cherchons à faire varier pour en déterminer l'impact.

2. L'annotation sémantique comme trait

L'objectif recherché à travers l'exploitation de l'annotation lexicale sémantique est une plus grande généricité dans la représentation des documents. Comme nous l'avons vu, la taxonomie permet de rassembler sous une même annotation des termes distincts renvoyant à un même concept. En utilisant cette annotation en tant que trait pour la méthode statistique, on permet à l'algorithme d'identifier une certaine communauté de sens entre des documents présentant des termes différents sous forme de chaînes de caractères, mais semblables dans leur position au sein de la représentation du domaine qu'est la taxonomie. Par exemple, dans les phrases "J'ai rencontré un conseiller très poli" et "Le technicien était incompétent", l'annotation SOCIETE#ENTITE#INTERLOCUTEUR sur "conseiller" et "technicien" indique que ces deux termes renvoient au même concept du domaine de la relation client. La hiérarchisation permet également de rapprocher des concepts appartenant à une même branche. Si l'on considère la phrase "Votre enseigne pratique des tarifs trop élevés", l'annotation "SOCIETE#ENTITE#STRUCTURE" sur "enseigne" indique que ce terme a également une certaine proximité avec les deux précédents. Au sein de notre système, la hiérarchie des concepts de la taxonomie est gérée de la façon suivante : chaque branche de la taxonomie constitue une valeur de trait. Ainsi, "conseiller" et "technicien" ont trois traits en commun, et deux avec "enseigne".

Lors de la phase de classification, la sortie du classifieur est, pour chaque catégorie, une probabilité d'appartenance du document à celle-ci, que l'on appelle le score qui est donc compris entre 0 et 1.

C Méthode symbolique

1. Tokensregex

Constituée de règles établies par un linguiste, la méthode symbolique a pour but de corriger les erreurs et d'améliorer les résultats de la méthode statistique. Le formalisme utilisé dans notre système est celui des tokensregex [Chang et Manning, 2014], également développé par le Natural Language Processing Group de l'université de Stanford. Il s'agit d'une syntaxe d'expressions régulières permettant de détecter des patterns à base de tokens, et non simplement de chaînes de caractères comme c'est le cas des expressions régulières classiques. De plus, les tokensregex présentent une correspondance des expressions textuelles recherchées sous forme d'objet Java, ce qui permet une intégration aisée au sein de la plateforme de notre système de classification. Les règles de type tokensregex se composent d'un pattern et d'une action à exécuter lorsque le pattern est détecté dans le document. Dans notre cas, l'action consiste à agir sur le score attribué par la méthode statistique à une catégorie pour un document. Une action de type "BOOST" ajoute 1 au score, ce qui représente un impact très fort sur la classification.

2. Recherche de pattern

Les patterns portent sur les tokens sur lesquels on peut travailler à travers les différentes caractéristiques issues de la représentation des documents. Pour chaque token, nous disposons des informations suivantes :

- word : forme du token tel qu'il apparait dans le texte
- lemma : lemma du token
- part-of-speech : catégorie syntaxique du token
- annotation sémantique basée sur la taxonomie

A cela peuvent s'ajouter des caractéristiques comme le genre, la personne, le nombre, la place du token dans la phrase (première ou dernière position), etc., qui pour l'instant ne sont pas exploitées dans notre système.

Enfin, l'annotation lexicale sémantique issue de la taxonomie représente pour nous une information centrale dans l'élaboration de nos règles symboliques.

Chaque pattern peut porter sur une ou plusieurs de ces caractéristiques. De plus, les opérateurs booléens et la syntaxe habituelle des expressions régulières (quantifieurs, ensemble de caractères, références et priorités) sont implémentés dans ce formalisme. L'exemple suivant montre une règle détectant une phrase contenant :

- un terme désignant un interlocuteur de la société répertorié dans la taxonomie (annotation sémantique lexicale `SOCIETE#ENTITE#INTERLOCUTEUR ;`
- suivi de zéro à trois mots ;
- suivi du token " différent " sous une de ses différentes formes.

```
{ pattern: (([{sa:/SOCIETE#ENTITE#INTERLOCUTEUR.*/}]) /.+/{0,3}
            [{word:/diff[éeè]rent?e?s?/}]),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:6") ),
  name: "CRM6A" }
```

Figure 10 : Exemple de tokensregex

On voit ici que la généricité issue de la taxonomie peut être exploitée pour l'écriture des règles. En effet, grâce à l'annotation lexicale sémantique, il est possible de détecter un pattern portant sur un concept (interlocuteur de la société), et ayant des représentations textuelles différentes (conseiller, opératrice, etc.). L'objectif de cette règle est de classer dans la catégorie 6#STABILITE les verbatims parlant d'interlocuteurs différents.

D Amélioration du système

1. Performances et pistes d'évolution

Notre système relève d'une méthode hybride, c'est-à-dire combinant statistique et symbolique. Les performances d'un classifieur se mesurent en terme de précision et de rappel que nous cherchons à améliorer en développant la représentation des documents et la méthode symbolique.

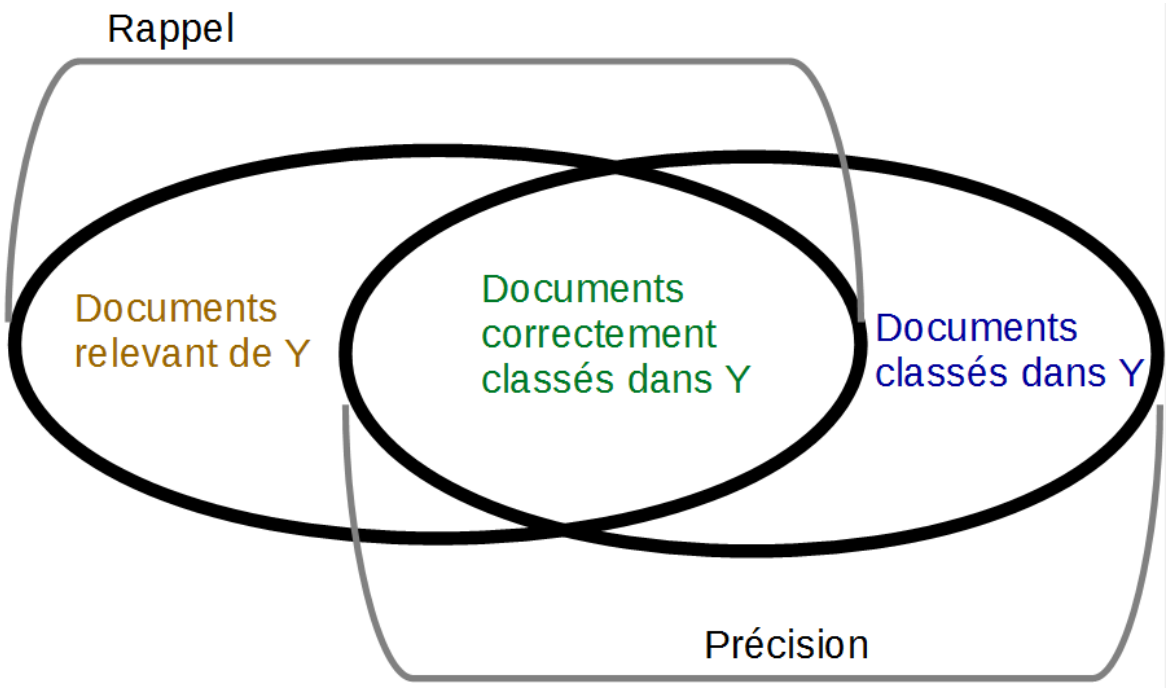


Figure 11 : Précision et rappel pour mesurer les performances d'un classifieur

Une faible précision indique que le système produit du bruit : il classe dans une catégorie des documents qui n'en relèvent pas. Cela étant notamment généré par des problèmes d'ambiguïté du lexique, il convient de travailler sur la représentation des documents. Il serait également possible de mettre en place des règles empêchant la classification dans telle ou telle catégorie, de documents présentant certains patterns. Cependant, cela nous semble à ce stade trop risqué : de telles règles peuvent être très restrictives et impacter alors négativement le rappel.

Un rappel faible se manifeste par du silence : une partie des documents relevant d'une catégorie n'est pas classée dans celle-ci. Ici encore, une meilleure représentation des documents peut diminuer le phénomène. De plus, les règles de classification permettent d'identifier des documents présentant des patterns particuliers, et donc de les classer dans la catégorie adéquate.

2. Cycle de développement

Afin de travailler précisément au développement du système, nous mettons en place un cycle permettant de se baser sur les résultats, et en particulier les erreurs constatées dans la classification, pour l'amélioration de la représentation des documents, c'est-à-dire de la taxonomie, et des règles de la méthode symbolique.

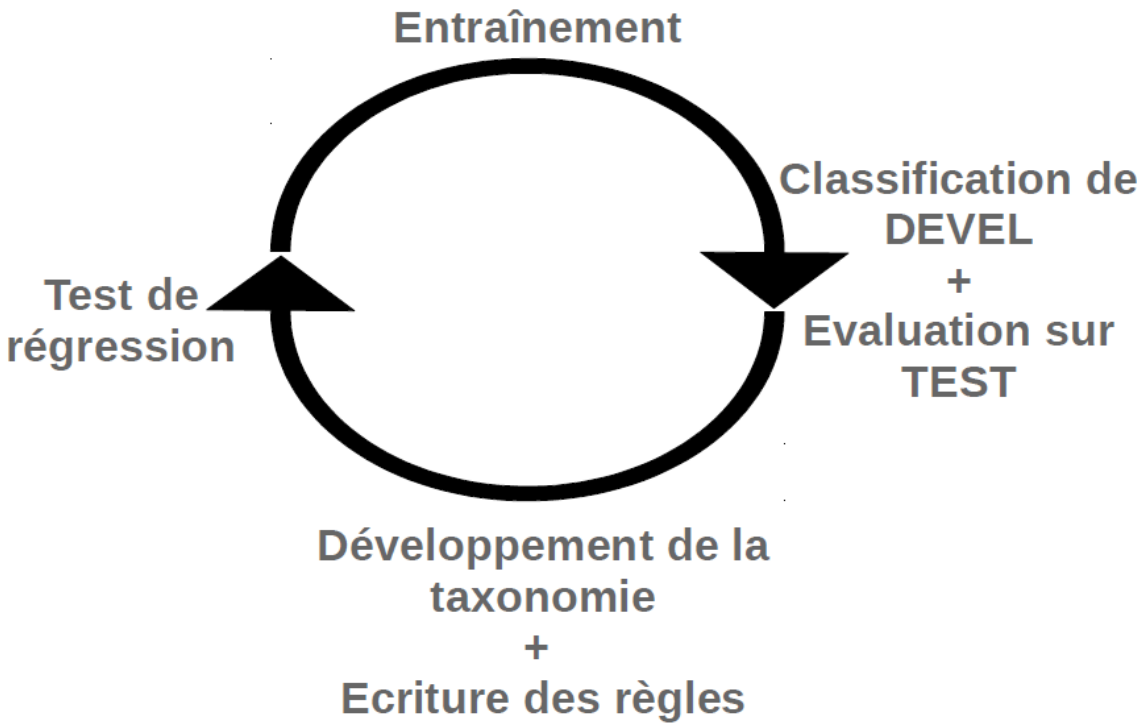


Figure 12 : Cycle de développement

a Entrainement

La première étape est la phase d'entrainement de l'algorithme. A partir des données annotées du corpus TRAIN, un modèle de classification est établi. Lors du premier cycle, la taxonomie construite à partir de l'extraction automatique des concepts est exploitée pour la représentation des documents. Lors des cycles suivants, s'y ajoutent de nouvelles entrées de la taxonomie.

b Classification et évaluation

Une fois l'algorithme entrainé, une classification du corpus DEVEL est effectuée, avec, pour la représentation des documents, la même taxonomie qu'à l'étape d'entraînement.

En parallèle, nous mesurons les performances du système sur le corpus TEST (le processus d'évaluation détaillé et les mesures utilisées sont présentées dans la troisième partie). L'objectif de cette évaluation est de suivre les progrès au fur et à mesure des cycles. De plus, nous disposons de résultats par catégorie afin de nous guider dans nos développements en identifiant les classes qui présentent le plus de problèmes. Pour obtenir ces données, deux classes Java (SentenceWithCategoriesList.java,

XmlToCsvClassification.java) ont été spécifiquement développées (cf. Annexe 1 : Développements informatiques réalisés).

c Développement de la taxonomie et écriture de règles

La troisième étape fait ici intervenir les compétences du linguiste computationnel, puisqu'il s'agit d'étoffer la taxonomie et d'écrire des règles de classification. Pour cela, nous nous basons sur une exploration du corpus DEVEL classifié. Un format de sortie des données a été spécialement conçu dans ce but : pour chaque verbatim, les scores (probabilité d'appartenance à une catégorie) attribués à chaque catégorie sont listés. Le but est d'identifier les erreurs de classification. Il faut noter ici que le corpus DEVEL n'étant pas annoté préalablement, il n'est pas possible de détecter de façon automatique les erreurs. On pourrait être tentés d'utiliser les corpus TRAIN ou TEST, mais comme nous l'avons vu, cela amènerait à un biais de "corpus over-fitting" et à une évaluation non objective (car l'amélioration du système aurait porté directement sur les cas du TEST et du TRAIN). Ainsi, il faut parcourir "manuellement" le corpus DEVEL et ses résultats pour rechercher des pistes d'amélioration.

Deux cas d'erreurs se présentent. Les faux positifs où le score pour une classe est anormalement élevé alors que le verbatim n'en relève pas, ou à l'inverse les faux négatifs quand le score est bas alors que la catégorie est pertinente pour ce document. L'amélioration du système porte concrètement sur deux éléments : les entrées de la taxonomie et les règles de classification. Pour résoudre des problèmes de faux négatifs, on cherche à ajouter des termes dans la taxonomie et des règles. Les faux positifs quant à eux doivent amener à repenser, voire à supprimer des éléments de la taxonomie ou des règles.

Pour chaque piste détectée, il est primordial d'en vérifier sa généricité sur le corpus. En effet, il ne serait pas pertinent de traiter des cas uniques ou peu représentés. Un autre écueil serait de travailler sur des termes ou patterns ambigus. Le terme "service" par exemple semble un bon candidat pour l’annotation SOCIETE#ENTITE #DEPARTEMENT, mais une exploration du corpus montre rapidement qu’il apparait bien souvent pour désigner les services offerts par la société, et non un département. Il est nécessaire dans ce cas de désambiguïser le terme grâce à son contexte : nous intégrons donc dans la taxonomie les termes "service client" et "service après-vente" qui sont fréquents dans le corpus.

Comme bien souvent en Traitement Automatique du Langage, la méthodologie de développement doit être guidée par les usages attestés en se basant sur un corpus représentatif, comme nous l'avons mis en place avec le corpus DEVEL. Il est à noter que nous élargissons à la taxonomie aux adjectifs et aux verbes qui nous semblent très pertinents pour la représentation du domaine.

Au cours de ce travail de développement, trois aspects sont à prendre en compte. Tout d'abord, l'impact très fort des règles (une règle ajoute 1 à un score de classification) incite à les utiliser avec parcimonie. Nous privilégions donc l'élargissement de la taxonomie qui agit sur la représentation des documents et recourons à une règle lorsque celle-ci s'avère indispensable : c'est le cas pour des expressions idiomatiques ou pour des catégories avec une représentation si faible que les données d'entrainement ne suffisent pas à établir un modèle solide dans la méthode statistique. De plus, une taxonomie riche permet de mieux représenter l'information sur le plan sémantique et d'obtenir ainsi un modèle d'apprentissage automatique plus solide.

D'autre part, il faut anticiper le développement de la taxonomie. L'ajout de nouveaux termes et la création de branches peuvent influencer fortement la classification à base de règles qui porteraient sur l'annotation sémantique lexicale. Une attention particulière doit donc être portée aux règles quand on travaille sur la taxonomie et vice versa, car les deux ressources s'impactent mutuellement.

Enfin, il faut trouver un équilibre entre la généricité et la spécificité des règles. En effet, une règle trop générale risque d'apporter du bruit dans les résultats. A l'inverse, on peut être tenté de mettre en place des règles très spécifiques, qui répondent au cas par cas, mais cela est très chronophage, et risque de ne pas être adapté pour de nouvelles entrées. Il faut donc toujours chercher à trouver le niveau de généricité le plus élevé qui génère le moins de bruit possible.

La taxonomie complète et les règles symboliques sont présentées respectivement en Annexe 2 : Taxonomie, et Annexe 3 : Règles symboliques.

d Test de régression

A la dernière étape, la vérification des résultats se fait par un test de régression : on compare les résultats de classement du DEVEL obtenus à l'issu de ce cycle, avec les résultats obtenus de la version antérieure. Ainsi, seules les différences sont mises en

évidence. Il est primordial de mettre en place ce type de tests pour vérifier que les évolutions apportées au système améliorent l'analyse et non l'inverse.

L'exemple suivant montre une règle générant du bruit, repérée dans le cadre d'un test de régression :

```
{ pattern: ( [{lemma:/répondre/}] /.+/{0,5} ([{sa:
/INFORMATION#QUESTION/}])),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:18") ),
  name: "CRM18B" }
```

Figure 13 : Exemple de tokensregex générant du bruit identifié grâce au test de régression

L'annotation INFORMATION#QUESTION renvoyant aux lemma "demande" et "question", ce pattern détecte des verbatims qui sont pertinents dans la catégorie 18#INFOS, tel que "le conseillé à répondu à mes questions", mais également des phrases qui n'en relèvent pas comme "ce produit ne répond pas à ma demande". La règle est donc modifiée et restreinte au lemma "question" :

```
{ pattern: ( [{lemma:/répondre/}] /.+/{0,5} ([{lemma:/question/}])),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:18") ),
  name: "CRM18B" }
```

Figure 14 : Exemple de tokensregex corrigée grâce au test de régression

Comme nous l'avons vu, le développement du système se fait par un processus d'itération des tâches suivantes : construction du modèle de classification, classification d'un corpus, correction et amélioration du système sur la base des erreurs repérées, test de régression. Une fois établie une version améliorée de notre classifieur, il convient d'en mesurer précisément les performances.

Partie 3

-

Evaluation du classifieur

L'évaluation de système de catégorisation par méthode hybride consiste à observer l'adéquation du classement qu'il génère avec le classement correct établi par un opérateur humain. Dans notre cas, nous comparons les résultats de la classification du corpus TEST avec l'annotation manuelle effectuée sur ce même corpus (décrite en partie 1). Notre objectif est d'évaluer l'apport des différentes ressources mises en œuvre, en particulier sur le plan sémantique. Nous souhaitons ainsi vérifier notre hypothèse, à savoir l'amélioration des performances d'un classifieur grâce à l'enrichissement sémantique.

A Méthodologie d'évaluation

1. Mesures utilisées et méthodes de calcul

Notre évaluation se base sur des mesures couramment utilisées dans la littérature (rappel, précision, F1, accuracy et Hammingloss) comme le présentent notamment [Tsoumakas et Vlahavas, 2007] et [Sorower, 2010]. Ces mesures manipulent quatre types de résultats :

- TP (true positive) : catégorie attribuée au document par le classifieur et par l'annotateur ;
- FP (false positive) : catégorie attribuée au document par le classifieur mais pas par l'annotateur ;
- TN (true negative) : catégorie attribuée au document ni par le classifieur ni par l'annotateur ;
- FN (false negative) : catégorie non attribuée au document par le classifieur mais attribuée par l'annotateur.

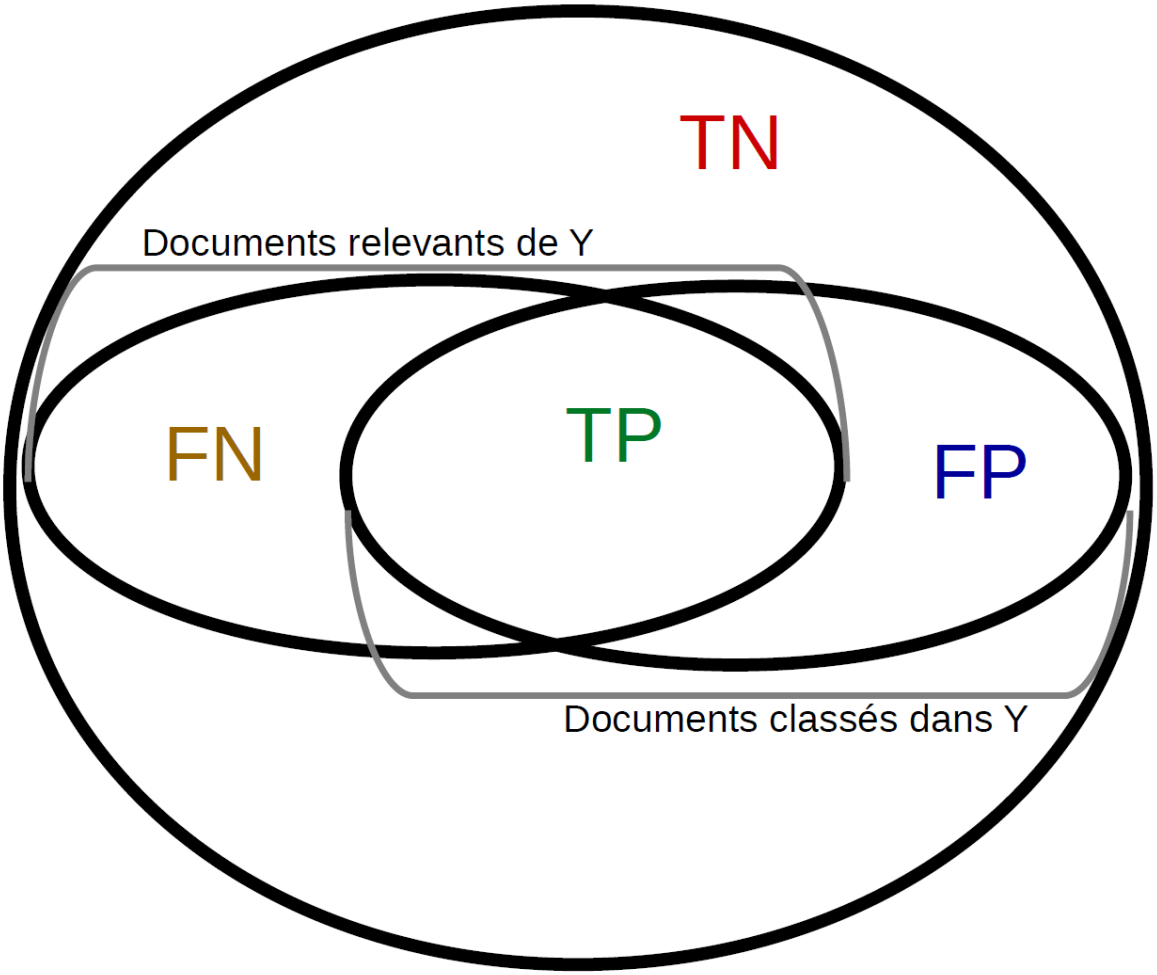


Figure 15 : Quatre types de résultats pour l'évaluation

Le nombre de chacun d'eux (TP, FP, TN, FN) permet de calculer :

- le rappel R (déjà évoqué dans la partie 2) : nombre de documents bien classés dans la catégorie Y par rapport au nombre de documents appartenant à la catégorie Y, formule : TP/(TP+FN) ;
- la précision P (également en partie 2) : nombre de documents bien classés dans la catégorie Y par rapport au nombre de documents classés la catégorie Y, formule : TP/(TP+FP) ;
- F1 : combine les deux mesures précédentes, formule : 2xPxR/(P+R) ;
- l'accuracy A : nombre de documents bien classés, formule : TP/(TP+FP+FN) ;
- Hammingloss H, formule : (FP+FN)/Nb total de catégories.

Il est à noter que notre méthodologie d'évaluation doit correspondre à un cadre de classification multicategorielle. En effet, un verbatim peut relever de plusieurs catégories et l'on attend du classifieur qu'il les identifie toutes. Il ne s'agit donc pas seulement de vérifier si pour un verbatim, sa catégorie est détectée, mais de vérifier si chacune des catégories dont il relève est attribuée, et pas d'autres. Par exemple, pour un verbatim Y, si l'annotation donne les catégories A et B, et le classifieur les catégories A, C et D, nous constatons différents phénomènes : le nombre de catégorie n'est pas le même, et celles attribués dans les deux cas ne sont que partiellement semblables.

[Tsoumakas et Vlahavas, 2007] décrivent précisément ce problème et les méthodes de calcul qui s'y associent. Deux approches sont proposées : l'approche par les documents et l'approche par les catégories. Celles-ci sont largement utilisées dans la littérature relative aux problèmes de classification multi-catégorielle.

La première consiste à faire une évaluation par document, comme le montre l'exemple suivant :

Catégories attribuées au document 1 par l'annotateur	Catégories attribuées au document 1 par le classifieur
Categorie A ->FN	Categorie B -> TP
Categorie B -> TP	Categorie D -> FP
Categorie C -> FN	

On obtient alors : 1 TP (B), 1 FP (D), 2 FN (A et C). Les TN sont les éventuelles autres catégories qui n'apparaissent pas dans ce tableau.

A partir de ces données, il est possible de calculer les cinq mesures citées précédemment, puis d'en faire une moyenne sur l'ensemble des documents.

Dans la seconde approche, une évaluation par catégorie est réalisée ainsi :

Documents attribués à la catégorie Y par l'annotateur	Documents attribués à la catégorie Y par le classifieur
Document 1 -> FN	Document 2 -> TP
Document 2 -> TP	Document 3 -> FP
	Document 4 -> FN

Les résultats ici sont : 1 TP (Document 2), 2 FP (Documents 3 et 4), 1 FN (Document) 1. Les TN sont les éventuels autres documents qui n'apparaissent pas dans ce tableau.

Une fois obtenus les TP, FP, TN, FN, pour toutes les catégories, deux calculs peuvent être effectués :

- la micro-moyenne : on établit la moyenne des TP, FP, TN, FN sur l'ensemble des catégories puis on calcule nos mesures d'évaluation (rappel, précision, F1, accuracy) ;
- la macro-moyenne : on calcule nos mesures d'évaluation (rappel, précision, F1 et accuracy) pour chacune des catégories, puis on en fait la moyenne.

Dans le cadre de cette approche par les catégories, nous n'utilisons pas la mesure Hammingloss qui n'a ici pas lieu d'être puisqu'elle est par nature une mesure d'approche par les documents.

2. Différentes configurations d'évaluation

A travers l'évaluation du système nous cherchons à déterminer les performances du classifieur, et en particulier à identifier l'intérêt des ressources que nous développons pour l'améliorer, à savoir la taxonomie et les règles symboliques. Il s'agit de vérifier si l'enrichissement sémantique permet au classifieur de générer de meilleurs résultats conformément à notre hypothèse de départ.

Pour cela nous dressons quatre configurations de classification mettant en œuvre nos différentes ressources. L'entrainement sur le corpus TRAIN puis l'évaluation sur le

corpus TEST vont nous permettre de comparer les résultats obtenus dans la section suivante. Les configurations se présentent de la manière suivante :

	BASIC	**TAXO_1**	**TAXO_2**	**SYMBO**
Représentation des documents	Word Lemma POS Dépendances syntaxiques Groupes nominaux	Word Lemma POS Dépendances syntaxiques Groupes nominaux Taxonomie semi-automatique	Word Lemma POS Dépendances syntaxiques Groupes nominaux Taxonomie semi-automatique Développement manuel de la taxonomie	Word Lemma POS Dépendances syntaxiques Groupes nominaux Taxonomie semi-automatique Développement manuel de la taxonomie
Méthode statistique	Apprentissage automatique sur corpus TRAIN	Apprentissage automatique sur corpus TRAIN	Apprentissage automatique sur corpus TRAIN	Apprentissage automatique sur corpus TRAIN
Méthode symbolique	Ø	Ø	Ø	Règles symboliques

A travers ces quatre configurations, une ressource est ajoutée à chaque étape. BASIC constitue notre "baseline", c'est-à-dire la configuration de base de départ, sans enrichissement sémantique, qui sera notre référence pour la comparaison des résultats avec les autres configurations. Chaque autre configuration implique une ressource supplémentaire à la précédente. TAXO_1 comporte la première version de la taxonomie décrite en partie 1 (issue de l'extraction automatique des concepts, 148 lemmes), TAXO_2 la version développée manuellement de la taxonomie décrite en partie 2 (126 nouveaux

lemmes), et SYMBO les règles symboliques également décrites en partie 2 (17 tokensregex).

3. Deux paramètres à fixer : seuil d'attribution et nombre de catégories

Comme nous l'avons vu, le principe du système est d'attribuer pour un verbatim, une probabilité d'appartenance (le score) pour chacune de nos 21 catégories. Ceci ne constitue pas en soi une catégorisation. Il s'agit donc de fixer deux paramètres afin de déterminer l'appartenance d'un verbatim à telle ou telle catégorie : le seuil de score au-delà duquel la catégorie doit être attribuée, et le nombre maximum de catégories (dépassant ce seuil) à retenir.

La valeur du seuil ne doit pas être fixée de façon arbitraire. Dans cette optique, avec les quatre configurations, nous procédons à différentes évaluations du classifieur avec des valeurs entre 0 et 1. Une démarche itérative nous montre que les meilleures performances sont obtenues avec un seuil se situant dans une fourchette allant de 0.08 à 0.2 (cf. Annexe 5 : Seuil d’attribution de catégorie).

D'autre part, le nombre maximum de catégories à retenir influe sur les résultats. En effet, on constate que dans le cas où un nombre important de catégories dépassent le seuil d'attribution, toutes ne sont pas pertinentes pour la catégorisation. Là encore, plusieurs évaluations faisant varier ce paramètre montrent que la valeur apportant les meilleurs résultats est 2. Ceci est à mettre en regard de l'annotation manuelle (décrite dans la première partie). [Tsoumakas et Katakis, 2008] proposent une mesure à ce sujet : "label cardinality" qui donne la moyenne du nombre de catégories attribuées par document. Sur le corpus annoté manuellement, elle est de 1.33. Il y a donc cohérence entre ce paramètre fixé à 2 et ce qui est réalisé par un opérateur humain.

Les différentes évaluations pour fixer le seuil d'attribution et le nombre maximum de catégories présentent une constance parmi les quatre configurations, ce qui nous conforte dans le choix de ces valeurs.

B Résultats et intérêt de l'enrichissement sémantique

Notre méthodologie d'évaluation étant définie, nous pouvons désormais effectuer une analyse détaillée des résultats. Rappelons ici que notre classifieur a pour but de constituer une couche générique dans le domaine de la relation client. Comme nous l'avons vu, les ressources sémantiques développées ne contiennent aucun élément relatif à un

secteur d'activité, cette couche étant prévue dans une évolution ultérieure de notre système. De ce fait, les résultats obtenus ne présentent pas des valeurs élevées, mais ce qui nous intéresse est de dégager dans ces résultats l'apport de l'enrichissement sémantique.

1. *Résultats par configurations*

Le tableau suivant est une synthèse des résultats (les résultats détaillés sont présentés en Annexe 4 : Résultats détaillés), où sont croisés les quatre configurations et les différentes méthodes d'évaluation. Le seuil d'attribution est de 0.1 et le nombre de catégories maximum de 2, comme pour tous les tableaux de cette section.

	Approche par les documents					Approche par les catégories – micro-moyenne				Approche par les catégories – macro-moyenne			
	P	R	F1	A	H	P	R	F1	A	P	R	F1	A
BASIC	53,55%	53,40%	51,47%	45,73%	0,0679	49,51%	50,33%	49,92%	33,26%	46,68%	39,22%	39,88%	27,94%
TAXO_1	54,74%	55,85%	53,08%	47,03%	0,0652	51,78%	53,33%	52,55%	35,63%	48,70%	45,13%	44,20%	31,23%
TAXO_2	55,45%	56,67%	53,74%	47,67%	0,0639	52,77%	54,00%	53,38%	36,40%	48,17%	46,25%	45,09%	32,30%
SYMBO	56,16%	57,42%	54,57%	48,67%	0,0625	53,70%	55,67%	54,66%	37,61%	49,52%	47,95%	46,67%	33,51%

Figure 16 : Comparaison des résultats sur les quatre configurations

La première remarque que l'on peut faire, générale, est que les résultats s'améliorent à chaque configuration, bien que cette évolution soit de faible ampleur. Cela montre que l'exploitation de ressources linguistiques pour l'enrichissement sémantique a effectivement un impact positif sur la catégorisation réalisée par notre système.

Si l'on regarde en particulier la mesure F1 qui prend en compte le rappel et la précision, on constate que la plus forte amélioration des performances se trouve à la première étape, c'est-à-dire au passage entre BASIC et TAXO_1 (avec la taxonomie semi-automatique de 148 lemmes). On peut supposer que cela dépend également fortement du nombre et du choix des lemmes de la taxonomie et du nombre de règles de la méthode symbolique : TAXO_2 comporte 126 nouveaux lemmes et SYMBO 17 règles. Il sera pertinent d'étudier à nouveau les résultats après une forte augmentation de ces éléments.

L'évolution du rappel et de la précision présente une tendance intéressante : dans chacune des trois méthodes de calcul, on observe que le rappel a une plus forte augmentation que la précision. Cela est en accord avec nos remarques évoquées dans le cadre de la méthodologie de développement de la méthode symbolique (partie 2) : le rappel, sur lequel on peut travailler en ajoutant des entrées à la taxonomie et des règles de

classification pour détecter plus de documents pertinents, est plus facile à améliorer que la précision, qui doit au contraire exclure des documents relevés par la méthode statistique. Or, comme nous l'avons évoqué, l'effet de règles symboliques d'exclusion est difficile à anticiper et présente un risque de diminution du rappel, c'est pourquoi nous n'en intégrons pas dans le système à l'heure actuelle.

Globalement, ces résultats montrent que notre hypothèse d'amélioration d'un classifieur par enrichissement sémantique se vérifie. La taxonomie, qu'elle soit générée semi-automatiquement ou manuellement, ainsi que les règles symboliques, apportent chacune de meilleurs résultats. Elles permettent ainsi de combler les lacunes de la méthode statistique dont le modèle n'est pas assez robuste du fait de la petite taille du corpus d'entrainement (un des verrous évoqués en introduction).

2. Résultats par catégories

Afin de comparer les performances du classifieur avec ou sans enrichissement sémantique par catégories, nous dressons le tableau suivant qui présente la mesure F1 de la première et de la dernière configuration (BASIC et SYMBO). Pour rappel, cette mesure est calculée pour chaque catégorie sur la base des TP, FP et FN. Nous donnons également le nombre de phrases annotées (par l'opérateur humain) et le nombre de phrases catégorisées (par le système) pour mettre en regard les résultats avec de la taille du corpus sur chaque catégorie.

		1#ACCUEIL	2#ATTITUDE	3#CONFORT	4#ATTENTE	5#ACCESSIBILITE	6#STABILITE	7#ASSISTANCE
	Nb phrases annotées	14	22	9	18	15	3	4
BASIC	Nb phrases catégorisées	13	28	10	16	13	0	5
	F1	88,89%	52,00%	73,68%	70,59%	50,00%	0,00%	0,00%
SYMBO	Nb phrases catégorisées	12	26	9	17	18	7	4
	F1	92,31%	66,67%	66,67%	74,29%	60,61%	40,00%	0,00%

		8#PROACTIVITE	9#COMPREHENSION	10#COMPETENCE	11#FIDELITE	12#DEPART	13#ENGAGEMENT	14#OFFRE
	Nb phrases annotées	4	7	46	5	7	5	18
BASIC	Nb phrases catégorisées	0	2	52	7	6	2	30
	F1	0,00%	44,44%	51,02%	16,67%	46,15%	28,57%	45,83%
SYMBO	Nb phrases catégorisées	1	3	45	6	8	2	31
	F1	0,00%	60,00%	57,14%	36,36%	40,00%	28,57%	44,90%

		15#TARIFS	16#ACHAT	17#CONTRAT	18#INFOS	19#IMAGE	20#CONCURRENCE	21#HORSCATEG
	Nb phrases annotées	31	18	5	9	10	12	38
BASIC	Nb phrases catégorisées	26	22	6	4	3	2	58
	F1	66,67%	45,00%	36,36%	46,15%	15,38%	14,29%	45,83%
SYMBO	Nb phrases catégorisées	30	21	7	7	6	4	47
	F1	72,13%	41,03%	50,00%	62,50%	12,50%	25,00%	49,41%

Figure 17 : Comparaison des résultats sur les 21 catégories

Dans un premier temps, nous pouvons remarquer qu'il y a une répartition très inégale du nombre de phrases annotées (par l'opérateur humain) ou catégorisée (par le système) parmi les vingt-et-une catégories. Lorsque ce nombre est important, les résultats sont relativement cohérents avec ceux de l'ensemble du corpus. C'est le cas par exemple sur la catégorie 10#COMPETENCE (46 phrases annotées) dont la mesure F1 est de 51,02% dans la configuration BASIC, et de 57,14% dans la configuration SYMBO. D'autres catégories, pour lesquelles il y a peu, voire très peu de phrases annotées, présentent des résultats médiocres, comme par exemple 7#ASSISTANCE (4 phrases annotées, F1 à 0% avec ou sans enrichissement sémantique).

On voit ici tout l'impact du corpus d'entrainement. Si celui-ci est trop petit, il ne permet pas d'établir un modèle solide, seule la méthode symbolique permet alors d'obtenir de meilleurs résultats : 6#STABILITE passe d'une F1 à 0% en BASIC à 40% en SYMBO grâce à l'application d'une règle symbolique, tandis que 8#PROACTIVITE qui n'a pas de règle reste à 0%.

D'autre part, on peut se demander pourquoi certaines catégories ont de meilleurs résultats que d'autres pour un nombre de phrases annotées du même ordre : 19#IMAGE (10 phrases annotées, une règle symbolique) et 20#CONCURRENCE (12 phrases annotées, 2 règles symboliques) ont une F1 entre 12,5% et 25% seulement, tandis que 3#CONFORT (9 phrases, pas de règle) et 1#ACCUEIL (14 phrases et pas de règle) ont obtenu des F1, entre 66% et 92%. Ici, on peut supposer que le type de phrase n'est pas le même et peut être difficile à gérer en terme de représentation des documents. 19#IMAGE peut se manifester à travers une grand variété d'expressions et de l'implicite : "quelle honte !", "la délocalisation n'est pas une bonne chose". 20#CONCURRENCE quant à elle est fortement dépendante des noms propres des concurrents en question, qui ne sont pas présents dans notre taxonomie puisque celle-ci est générique (mais ce problème devrait être levé lors du développement d'une couche spécifique). A l'inverse, 3#CONFORT et 1#ACCUEIL semblent plus simples à représenter, grâce à un vocabulaire récurrent, comme par exemple "accueil" qui se trouve dans la taxonomie.

Enfin, concernant l'apport de l'enrichissement sémantique, il se manifeste par un gain de performances sur 15 catégories. Lorsque ce n'est pas le cas, cela concerne des catégories avec peu de phrases annotées et de règles dans la méthode symbolique, et rejoint notre première remarque : le modèle statistique ne peut donner de bons résultats avec un

corpus d'entrainement trop petit. Seules les règles symboliques peuvent alors avoir un effet positif.

Globalement, nous notons ici les limites d'une évaluation portant sur un corpus de très petite taille, pour lequel il est difficile de généraliser les tendances observées. Celles-ci devront être confirmées par des expérimentations futures, mais nous notons d'ores-et-déjà qu'il ressort de cette analyse l'apport positif de l'enrichissement sémantique.

3. Résultats par corpus

Notre évaluation porte également sur la comparaison des résultats entre corpus comme le montre le tableau suivant qui contient la mesure F1 des trois méthodes de calcul (par les documents, par les catégories avec la micro-moyenne, puis la macro-moyenne) :

	BANQUE			MECA_AUTO		
	doc	cat_micro	cat_macro	doc	cat_micro	cat_macro
BASIC	40,80%	42,65%	24,42%	66,00%	64,66%	36,58%
SYMBO	48,13%	49,64%	32,64%	65,00%	65,19%	36,19%

	ASSURANCE			TELECOM		
	doc	cat_micro	cat_macro	doc	cat_micro	cat_macro
BASIC	44,10%	42,93%	29,61%	55,60%	52,41%	32,83%
SYMBO	47,60%	48,45%	39,76%	59,07%	57,93%	41,25%

Figure 18 : Comparaison des résultats sur les 4 corpus

Deux observations intéressantes se dégagent : le passage de BASIC à SYMBO améliore les résultats, sauf pour le corpus MECA_AUTO où ils stagnent. Cependant, il faut noter que sur ce corpus, la mesure F1 est, dès l'étape BASIC, au-dessus de celles des autres corpus. Ce phénomène tend à montrer que l'enrichissement sémantique est d'autant plus pertinent lorsque les résultats de la méthode statistique seule sont faibles.

La remarque soulevée dans la partie 1, où nous avons mis en avant le fait que le corpus ASSURANCE est plus "riche" (en nombre de tokens par phrase, et phrase par verbatims), ne semble finalement pas avoir d'impact positif sur les résultats, car ce corpus se place seulement en troisième position sur quatre en terme de performance.

4. *Comparaison avec d'autres travaux*

Afin de situer nos résultats par rapport à d'autres expérimentations, nous avons choisi les comparer avec ceux de [Hernandez, Jadi, Lark, et Monceaux, 2015] dans le cadre de la 11ème édition du DEFI Fouille de Textes, qui nous semble être, parmi la littérature évoquée dans l'état de l'art, relativement proche de notre cadre de travail.

Les expérimentations décrites dans [Hernandez, Jadi, Lark, et Monceaux, 2015] présentent un certain nombre de points communs avec notre travail : il s'agit de répondre à une tâche de classification de textes courts en 19 catégories, avec un petit corpus d'entrainement (et notamment une distribution déséquilibrée entre les classes dans celui-ci). Dans ce but, les auteurs ont mis en place un système d'apprentissage supervisé. A cela s'ajoute l'exploitation de lexiques pour l'enrichissement sémantique des documents (en l'occurrence des tweets). La différence notable entre la tâche décrite dans cet article et la nôtre est le type de catégories : tandis que nous travaillons sur les thèmes abordés dans nos textes, les classes de [Hernandez, Jadi, Lark, et Monceaux, 2015] portent sur des sentiments et opinions. D'autre part, leur classifieur donne pour chaque texte une seule catégorie en sortie et n'est donc pas multi-catégoriel comme le nôtre. C'est pourquoi nous ne sommes pas en mesure de comparer les résultats en termes de valeurs des mesures effectuées, mais nous nous penchons plutôt sur l'analyse des meilleures configurations du classifieur (avec ou sans recours à des lexiques). Le tableau suivant présente un récapitulatif de ces résultats :

	Micro-précision	Macro-précision	Macro-moyenne F1
Méthode Statistique sans lexique	27,92%	2,17%	28,50%
MS + lexiques endogène et exogène	33,43%	2,81%	30,00%
MS + lexique endogène seul	31,59%	2,73%	31,50%

Figure 19 : Récapitulatif des résultats de [Hernandez, Jadi, Lark, et Monceaux, 2015]

Ainsi, il nous semble pertinent d'établir quelques parallèles entre nos résultats respectifs. Le tableau 2 de [Hernandez, Jadi, Lark, et Monceaux, 2015] présente les résultats en terme de précision sur l'ensemble du corpus de test, avec et sans l'utilisation des lexiques. Leurs observations concordent avec les nôtres : l'apport de lexiques pour l'enrichissement sémantique des documents à classer améliorent les performances du système. A partir de leurs résultats par catégories des tableaux 5, 8 et 11, nous avons

également calculé la macro-moyenne de la mesure F1 pour chacune de leurs configurations. Cette évaluation confirme à nouveau l'intérêt de l'exploitation des lexiques. De plus, l'analyse détaillée des résultats par catégories indique, comme dans nos remarques dans la section précédente, que les classes ayant une très faible représentation dans le corpus d'entrainement n'obtiennent pas de bons résultats en phase de test. Cette observation est donc en faveur d'un développement de méthode symbolique de classification lorsque la méthode statistique n'est pas suffisamment performante.

Conclusion

A Bilan et acquis

Nous avons développé un système de classification multi-catégoriel de textes courts, fondé sur une méthode hybride, afin de déterminer quel est l'apport de l'enrichissement sémantique sur une telle tâche, avec peu de données d'entrainement. Nos résultats tendent à montrer que cet apport est positif car il améliore les performances du classifieur. L'annotation sémantique lexicale permet une meilleure représentation des documents, sur laquelle se basent les méthodes statistique et hybride. D'autre part, les règles symboliques comblent les lacunes de l'apprentissage automatique sur les catégories les moins représentées dans le corpus d'entrainement et dont le modèle statistique de classification n'est pas satisfaisant. Les différentes phases de ce travail nous ont permis d'acquérir une méthodologie et une expertise sur plusieurs points.

Tout d'abord, la définition des catégories, en recherchant les primitives du domaine, a impliqué la mise en œuvre d’une méthodologie rigoureuse afin d’aboutir à une liste répondant au mieux à notre tâche.

D’autre part, il faut noter que la préparation des corpus est une étape cruciale. En définissant précisément les besoins, on procède à une sélection et une mise en forme adéquate des documents. L'ensemble du système reposant sur ces données, il est impératif qu'elles soient représentatives de la tâche à réaliser, et exploitables grâce à une structure réfléchie.

Les corpus ont également été la source de l'élaboration de la taxonomie. A cette occasion, nous nous sommes approprié une méthodologie d'exploitation de lexique et de structuration rigoureuse qui pourra être réitérée en fonction de nouveaux besoins.

Par la suite, l'acquisition du formalisme des tokensregex nous a permis de développer la méthode symbolique de classification. Pour que celle-ci soit efficace et réponde aux lacunes de la méthode statistique, il était indispensable de mettre en place un cycle de développement raisonné.

Enfin, nous avons mis au point une évaluation des performances, complexe car il s'agit d'un système multi-catégoriel, et précise, grâce à laquelle il est possible de suivre les évolutions du classifieur, mais également de le comparer à d'autres systèmes.

B Limites et perspectives

Au cours de ce travail, nous avons été confrontés à un certain nombre de limites. Certaines ont été levées car elles nous semblaient être des priorités à résoudre (comme par exemple la représentation des documents par l'exploitation du lexique, ou encore l'évaluation précise des performances). D'autres, qui ne constituaient pas de verrous directs pour notre travail, ont été prises en compte et nous permettent d'émettre des perspectives à différents niveaux.

Les documents que nous analysons subissent des pré-traitements indispensables avant d'entrer dans la phase de représentation (analyse syntaxique, lexicale etc.). Il s'agit principalement de corrections orthographiques. Cela est d'autant plus important pour des textes comme ceux qui nous intéressent : orthographes, ponctuations, abréviations spécifiques SMS/web. Nous avons fait le choix d'utiliser tels quels les pré-traitements déjà à disposition dans la plateforme Holmes, et développés pour des projets antérieurs. Cependant, il nous paraitrait judicieux de tester différentes configurations de pré-traitements car ils peuvent avoir une influence sur les performances du classifieur comme l'indique [Abdaoui et al., 2015].

La représentation des documents, qui a été un point clef de notre travail, peut encore être développée et améliorée. Nous pensons notamment à l'ajout de lexique exogène comme nous l'avons vu dans [Hernandez, Jadi, Lark, et Monceaux, 2015]. Une voie plus ambitieuse encore serait de passer de la structure de taxonomie à une structure d'ontologie

où des relations sémantiques plus variées entre les concepts seraient intégrées, comme le décrit [Bachimont, 2000].

La méthode statistique, dont nous avons vu les limites causées par la petite taille du corpus d'entrainement, présente tout de même quelques pistes d'amélioration. En explorant les résultats document par document, nous avons constaté que le modèle a tendance à privilégier fortement (avec un score élevé) une seule catégorie, même si plusieurs sont effectivement pertinentes. Il conviendrait donc de chercher à paramétrer l'algorithme afin qu'il s'oriente plus vers une classification multi-catégorielle. D'autre part, nous avons fait le choix, par soucis de simplicité dans un premier temps, de fixer un seuil d'attribution de catégorie unique pour toutes les catégories. Toutefois, il serait bon d'envisager de faire évoluer le seuil en fonction des performances, catégorie par catégorie.

Enfin, l'évaluation, comme nous l'avons vu, est limitée par la très petite taille du corpus de test. Pour pallier à ce problème, deux pistes nous semblent intéressantes. La validation croisée que l'on retrouve régulièrement dans la littérature, et notamment [Hernandez, Jadi, Lark, et Monceaux, 2015] permet de consolider les résultats. D'autre part, une comparaison des mesures avec une attribution aléatoire des catégories sur le même corpus de test serait un bon indicateur.

Pour conclure, dans un tel travail, le nombre de paramètres sur lesquels on peut agir pour faire varier les performances est important, et par conséquent les différentes configurations du système possibles ne sont pas toutes explorables. Il s’agit donc de définir les priorités, en se focalisant sur les points qui paraissent a priori les plus pertinents, puis de les valider a posteriori par des expérimentations scientifiques.

Bibliographie

Abdaoui, A., Tapi Nzali, M. D., Azé, J., Bringay, S., Lavergne, C., Mollevi, C., et Poncelet, P. (2015). ADVANSE : Analyse du sentiment, de l'opinion et de l'émotion sur des Tweets Français. Présenté à 22ème Traitement Automatique des Langues Naturelles, Caen, France.

Abeillé, A., Clément, L., et Toussenel, F. (2003). Building a Treebank for French. In A. Abeillé (Éd.), *Treebanks* (p. 165-187). Springer Netherlands.

Actes de la 11e Défi Fouille de Texte (DEFT'2015). (2015). Présenté à Défi Fouille de Texte, Caen, France.

Actes du neuvième DÉfi Fouille de Textes (DEFT13). (2013). Présenté à Défi Fouille de Texte, Les Sables-d'Olonne, France.

Bachimont, B. (2000). Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en ingénierie des connaissances. (chapitre 19). In *Ingénierie des connaissances : évolutions récentes et nouveaux défis* (p. 305-323). Paris: Eyrolles.

Baharudin, B., Lee, L. H., et Khan, K. (2010). A Review of Machine Learning Algorithms for Text-Documents Classification. *Journal of Advances in Information Technology*, 1(1), 4-20.

Basili, R., Cammisa, M., et Moschitti, A. (2005). Effective use of WordNet semantics via kernel-based learning. In Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL).

Chang, A. X., et Manning, C. D. (2014). *TokensRegex: Defining cascaded regular expressions over tokens* (No. Technical Report CSTR 2014-02). Department of Computer Science, Stanford University.

Charlet, J., Bachimont, B., et Troncy, R. (2004). Ontologies pour le web sémantique. *Revue I3*.

Chen, M., Jin, X., et Shen, D. (2011). Short Text Classification Improved by Learning Multi-granularity Topics. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* - Volume Volume Three (p. 1776–1781). Barcelona, Catalonia, Spain: AAAI Press.

Dini, L., Bittar, A., et Ruhlmann, M. (2013). Approches hybrides pour l'analyse de recettes de cuisine. Présenté à Défi Fouille de Texte (DEFT13), Les Sables-d'Olonne, France.

Doucy, G., et Massoussi, T. (2012). Sémantique inférentielle et compréhension des verbatim clients. Présenté à Congrès Mondial de Linguistique Française – CMLF 2012.

Eensoo, E., et Valette, M. (2012). Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments (Vol. 2, p. 367-374). Présenté à TALN 2012, GETALP-LIG.

Eensoo, E., et Valette, M. (2014). Sémantique textuelle et TAL : un exemple d'application à l'analyse des sentiments. In *DOCUMENTS, TEXTES, OEUVRES. Perspectives sémiotiques* (à paraitre).

Godbole, S., et Sarawagi, S. (2004). Discriminative Methods for Multi-labeled Classification. In H. Dai, R. Srikant, et C. Zhang (Éd.), *Advances in Knowledge Discovery and Data Mining* (p. 22-30). Springer Berlin Heidelberg.

Grivel, L. (2007). Maîtriser le processus de text mining dans le cadre d'applications d'intelligence économique, de gestion de la relation client ou de gestion de connaissances. Présenté à Veille Stratégique Scientifique et Technologique (VSST), Marrakech (Maroc).

Hernandez, N., Jadi, G., Lark, J., et Monceaux, L. (2015). Exploitation de lexiques pour la catégorisation fine d'émotions, de sentiments et d'opinions. In *Actes de la 11e Défi Fouille de Texte* (p. 51–60).

Huang, Y., Murphey, Y. L., et Ge, Y. (2013). Machine learning of engineering diagnostic knowledge from unstructured verbatim text descriptions. In *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)* (p. 46-52).

Janik, M. (2008). *Training-less ontology-based text categorization* (thèse de doctorat). University of Georgia.

Maurel, S., Curtoni, P., et Dini, L. (2008). A hybrid method for sentiment analysis. Présenté à Défi Fouille de Texte 2007 (DEFT'07).

Mondary, T., Després, S., Nazarenko, A., et Szulman, S. (2008). Construction d'ontologies à partir de textes : la phase de conceptualisation. In *19èmes Journées Francophones d'Ingénierie des Connaissances (IC 2008)* (p. 87-98). Nancy, France.

Poirier, D., Fessant, F., et Tellier, I. (2010). De la Classification d'Opinion à la Recommandation : l'Apport des Textes Communautaires. *TAL : traitement automatique des langues : revue semestrielle de l'ATALA*, 51(3), 19-46.

Poudat, C., Cleuziou, G., et Clavier, V. (2006). Catégorisation de textes en domaines et genres. *Document numérique*, Vol. 9(1), 61-76.

Salperwyck, C., et Lemaire, V. (2011). Impact de la taille de l'ensemble d'apprentissage: une étude empirique. Présenté à Extraction et Gestion des Connaissances (EGC), Brest, France.

Sclano, F., et Velardi, P. (2007). TermExtractor: a Web Application to Learn the Shared Terminology of Emergent Web Communities. In R. J. Gonçalves, J. P. Müller, K. Mertins, et M. Zelm (Éd.), *Enterprise Interoperability II* (p. 287-290). Springer London.

Sorower, M. S. (2010*). A literature survey on algorithms for multi-label learning*. Oregon State University, Corvallis.

Toussaint, Y. (2011, novembre 21). *Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances* (HDR). Université Henri Poincaré - Nancy I.

Tsoumakas, G., et Katakis, I. (2008). Multi-label classification: An overview. In *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*. IGI Global.

Tsoumakas, G., et Vlahavas, I. (2007). Random k-Labelsets: An Ensemble Method for Multilabel Classification (p. 406-417). Présenté à European Conference on Machine Learning, Springer Berlin Heidelberg.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit* (thèse de doctorat). Université de Toulouse II le Mirail.

Vernier, M., Monceaux, L., Daille, B., et Dubreil, E. (2009). Catégorisation des évaluations dans un corpus de blogs multi-domaine. *Revue des Nouvelles Technologies de l'Information*, 45--70.

Vincent, M., et Winterstein, G. (2013). Construction et exploitation d'un corpus français pour l'analyse de sentiment. Présenté à TALN-RÉCITAL, Les Sables d'Olonne, France.

Table des illustrations

Table des annexes

Annexe 1
Développements informatiques réalisés

Cette annexe présente les développements informatiques réalisés dans le cadre de ce travail. Ils ont permis de manipuler les données à traiter en les organisant et en les structurant. Pleinement intégrés à la plate-forme Holmes, ils ont été conçus de façon modulaire pour s'intégrer à d'autres projets.

Classes Java

XmlDocumentManaging.java

Au sein d'un dossier contenant différents fichiers XML de verbatims, cette classe effectue un redécoupage selon des nombres de verbatims passés en paramètres et donne en sortie des fichiers CSV contenant une phrase par ligne. L'objectif est d'obtenir des corpus de tailles déterminées prêts à être annotés.

AnnotationStat.java

Cette classe prend en entrée un fichier CSV de phrases annotées et un fichier contenant la liste des catégories. Elle renvoie en sortie un fichier CSV qui donne pour chaque catégorie le nombre de phrases annotées. L'objectif est de dresser des statistiques de distribution des catégories dans un corpus.

ConceptValues.java

Cette classe construit et renvoie un objet qui stocke un nombre d'occurrences, une somme de scores et un identifiant de corpus. Cet objet est lui-même une propriété de l'objet concept.

XmlConceptOutput.java

Cette classe crée une liste dédoublonnée de concepts associés à leurs données (issues de chaque corpus). La sortie est un fichier CSV contenant cette liste.

XmlConceptRetrieval.java

Cette classe parcourt un dossier contenant des verbatims enrichis, puis construit et renvoie un objet qui stocke les concepts de ces verbatims et leurs infos.

TaxonomyManaging.java

Cette classe effectue un tri alphabétique des entrées des fichiers gazetteers contenant la taxonomie. L'objectif est une meilleure lisibilité.

SentenceWithCategoriesList.java

Cette classe construit et renvoie un objet qui contient lui-même un objet Sentence, ainsi qu'une liste de catégories sysCategories attribuées par la méthode statistique.

XmlToCsvClassification.java

Cette classe prend en entrée la sortie XML de la classification et donne en sortie un CSV contenant par ligne : la phrase classifiée, le score obtenu dans chacune des 21 catégories (les scores sont ordonnés de la catégorie 1 à 21), les identifiants de la phrase, du document et du corpus.

Script bash

create_taxonomy.sh

Ce script prend en entrée les différentes sources de la taxonomie (gazetteers et tokenregex) et effectue un tri alphabétique de chaque entrée sur son chemin dans l'arborescence de la taxonomie. L'objectif est une meilleure lisibilité.

Annexe 2
Taxonomie

Cette annexe contient la liste complète des entrées de la taxonomie telle qu'elle apparait à la fin de ce travail. Le tri est alphabétique sur le chemin dans l'arborescence. Elle donne également le fichier dans lequel se trouve l'entrée (gazetteer ou tokensregex) et le lemme, sous forme d'expression régulière le cas échéant. Cette taxonomie a vocation à s'étendre au fur et à mesure du développement ultérieur du classifieur.

CHEMIN#DANS#LARBORESCENCE	fichier_source	lemme
ACCUEIL	crm_adj_lemma_manual	acc?u?ei?ll?((ie?s?)|(ante?s?))
ACCUEIL	crm_nouns_lemma	accueil
ARGENT	crm_nouns_lemma	argent
ARGENT	crm_nouns_lemma	coût
ARGENT	crm_nouns_lemma	frais
ARGENT	crm_nouns_lemma	montant
ARGENT	crm_nouns_lemma	somme
ARGENT#EURO	crm_nouns_lemma	€
ARGENT#EURO	crm_nouns_lemma	euro
ASSISTANCE#PROBLEME	crm_nouns_lemma	difficulté
ASSISTANCE#PROBLEME	crm_nouns_lemma	dommage
ASSISTANCE#PROBLEME	crm_nouns_lemma	erreur
ASSISTANCE#PROBLEME	crm_nouns_lemma	litige
ASSISTANCE#PROBLEME	crm_nouns_lemma	panne
ASSISTANCE#PROBLEME	crm_nouns_lemma	problème
ASSISTANCE#PROBLEME	crm_nouns_lemma	réclamation
ASSISTANCE#PROBLEME	crm_nouns_lemma	souci
ASSISTANCE#PROBLEME	crm_nouns_lemma	tort
ASSISTANCE#SOLUTION	crm_nouns_lemma	aide
ASSISTANCE#SOLUTION	crm_nouns_lemma	assistance
ASSISTANCE#SOLUTION	crm_nouns_lemma	dépannage
ASSISTANCE#SOLUTION	crm_nouns_lemma	réparation
ASSISTANCE#SOLUTION	crm_nouns_lemma	résolution
ASSISTANCE#SOLUTION	crm_nouns_lemma	solution
ASSISTANCE#SOLUTION	crm_verbs_lemma_manual	r[éeè]soudre
ATTENDRE	crm_verbs_lemma_manual	attendre
ATTENDRE	crm_verbs_lemma_manual	patienter
ATTENDRE	crm_verbs_lemma_manual	poireauter
ATTITUDE	crm_adj_lemma_manual	aimable
ATTITUDE	crm_adj_lemma_manual	avenant
ATTITUDE	crm_adj_lemma_manual	comp[éeè]tente?s?
ATTITUDE	crm_adj_lemma_manual	courtois
ATTITUDE	crm_adj_lemma_manual	sympa
ATTITUDE	crm_adj_lemma_manual	sympathique
ATTITUDE	crm_from_regex	prendre pour un con
ATTITUDE	crm_from_regex	raccrocher au nez
ATTITUDE	crm_from_regex	se moquer de moi/nous/du client

ATTITUDE	crm_nouns_lemma	amabilité
ATTITUDE	crm_nouns_lemma	écoute
ATTITUDE	crm_nouns_lemma	gentillesse
ATTITUDE	crm_nouns_lemma	patience
ATTITUDE	crm_nouns_lemma	sourire
ATTITUDE	crm_nouns_lemma	sympathie
ATTITUDE	crm_nouns_lemma_manual	sourire
CLIENT	crm_nouns_lemma	famille
CLIENT	crm_nouns_lemma	particulier
CLIENT#EN COURS	crm_adj_lemma_manual	client
CLIENT#EN COURS	crm_nouns_lemma	client
CLIENT#EN COURS	crm_nouns_lemma	clientèle
COMPETENCE	crm_adj_lemma_manual	capable
COMPETENCE	crm_nouns_lemma	compétence
COMPETENCE	crm_nouns_lemma	conseil
COMPETENCE	crm_nouns_lemma	efficacité
COMPETENCE	crm_nouns_lemma	expertise
COMPETENCE	crm_nouns_lemma	profess?ionn?alisme
COMPETENCE	crm_nouns_lemma	qualité
COMPETENCE	crm_nouns_lemma	rapidité
COMPETENCE	crm_nouns_lemma	réactivité
COMPETENCE	crm_nouns_lemma	responsabilité
COMPREHENSION	crm_nouns_lemma	compréhension
COMPREHENSION	crm_nouns_lemma_manual	incompréhension
COMPREHENSION#LINGUISTIQUE	crm_adj_lemma_manual	français
COMPREHENSION#LINGUISTIQUE	crm_nouns_lemma	accent
COMPREHENSION#LINGUISTIQUE	crm_nouns_lemma	étranger
COMPREHENSION#LINGUISTIQUE	crm_nouns_lemma	français
COMPREHENSION#LINGUISTIQUE	crm_nouns_lemma	langue
CONCURRENCE	crm_adj_lemma_manual	concurrent
CONCURRENCE	crm_nouns_lemma	concurr?ence
CONCURRENCE	crm_nouns_lemma	concurr?ent
CONTRAT	crm_adj_lemma_manual	souscrit
CONTRAT	crm_nouns_lemma	contrat
CONTRAT	crm_nouns_lemma	souscription
CONTRAT	crm_nouns_lemma_manual	souscription
CONTRAT	crm_verbs_lemma_manual	souscrire
DELAI	crm_adj_lemma_manual	long
DELAI	crm_nouns_lemma	attente
DELAI	crm_nouns_lemma	délai
DELAI	crm_nouns_lemma	échéance
DELAI	crm_nouns_lemma	retard
DELAI#TEMPS	crm_from_regex	sans cesse
DELAI#TEMPS	crm_nouns_lemma_manual	an
DELAI#TEMPS	crm_nouns_lemma_manual	année
DELAI#TEMPS	crm_nouns_lemma_manual	heure
DELAI#TEMPS	crm_nouns_lemma_manual	jour
DELAI#TEMPS	crm_nouns_lemma_manual	journée
DELAI#TEMPS	crm_nouns_lemma_manual	minute
DELAI#TEMPS	crm_nouns_lemma_manual	mois
DELAI#TEMPS	crm_nouns_lemma_manual	semaine
DEPART	crm_nouns_lemma	départ
DEPART	crm_nouns_lemma	résiliation
DEPART	crm_nouns_lemma_manual	rétractation
DEPART	crm_verbs_lemma_manual	désabonner
DEPART	crm_verbs_lemma_manual	fuir
DEPART	crm_verbs_lemma_manual	résilier
DEPART	crm_verbs_lemma_manual	rétracter
DEPART	crm_verbs_lemma_manual	rompre
DISPONIBILITE	crm_adj_lemma_manual	disponible

DISPONIBILITE	crm_adj_lemma_manual	dispos?
DISPONIBILITE	crm_adj_lemma_manual	joignable
DISPONIBILITE	crm_nouns_lemma	disponibilité
DISPONIBILITE	crm_nouns_lemma	distance
DISPONIBILITE	crm_nouns_lemma	horaire
DISPONIBILITE	crm_nouns_lemma	ouverture
DISPONIBILITE	crm_nouns_lemma	proximité
ENGAGEMENT	crm_nouns_lemma	engagement
ENGAGEMENT	crm_nouns_lemma	garantie
ENGAGEMENT	crm_nouns_lemma	promesse
ENGAGEMENT	crm_nouns_lemma_manual	parole
FIDELITE	crm_nouns_lemma	ancienneté
FIDELITE	crm_nouns_lemma	fid[eéè]lit[eéè]
FIDELITE	crm_nouns_lemma_manual	fid[éeè]lisation
FIDELITE	crm_verbs_lemma_manual	fid[éeè]liser
IMAGE	crm_adj_lemma_manual	honteux?
IMAGE	crm_nouns_lemma	confiance
IMAGE	crm_nouns_lemma	pub
IMAGE	crm_nouns_lemma	publicité
IMAGE	crm_nouns_lemma_manual	honte
INFORMATION#QUESTION	crm_nouns_lemma	demande
INFORMATION#QUESTION	crm_nouns_lemma	question
INFORMATION#REPONSE	crm_adj_lemma_manual	inform[éeè]e?s?
INFORMATION#REPONSE	crm_adj_lemma_manual	renseign[éeè]
INFORMATION#REPONSE	crm_adj_lemma_manual	renseign[éeè]e?s?
INFORMATION#REPONSE	crm_from_regex	raconter n'importe quoi
INFORMATION#REPONSE	crm_nouns_lemma	document
INFORMATION#REPONSE	crm_nouns_lemma	explication
INFORMATION#REPONSE	crm_nouns_lemma	information
INFORMATION#REPONSE	crm_nouns_lemma	questionnaire
INFORMATION#REPONSE	crm_nouns_lemma	renseignement
INFORMATION#REPONSE	crm_nouns_lemma	réponse
INTERACTION#CONTACT	crm_verbs_lemma_manual	contacter
INTERACTION#CONTACT	crm_verbs_lemma_manual	joindre
INTERACTION#CONVERSATION	crm_nouns_lemma	appel
INTERACTION#CONVERSATION	crm_nouns_lemma	communication
INTERACTION#CONVERSATION	crm_nouns_lemma	contact
INTERACTION#CONVERSATION	crm_nouns_lemma	conversation
INTERACTION#CONVERSATION	crm_nouns_lemma	rdv
INTERACTION#CONVERSATION	crm_nouns_lemma	rendez-vous
INTERACTION#CONVERSATION	crm_nouns_lemma	visite
INTERACTION#MEDIA	crm_nouns_lemma	courrier
INTERACTION#MEDIA	crm_nouns_lemma	envoi
INTERACTION#MEDIA	crm_nouns_lemma	internet
INTERACTION#MEDIA	crm_nouns_lemma	lettre
INTERACTION#MEDIA	crm_nouns_lemma	ligne
INTERACTION#MEDIA	crm_nouns_lemma	mail
INTERACTION#MEDIA	crm_nouns_lemma	message
INTERACTION#MEDIA	crm_nouns_lemma	recommandé
INTERACTION#MEDIA	crm_nouns_lemma	sms
OFFRE	crm_nouns_lemma	abonnement
OFFRE	crm_nouns_lemma	offre
OFFRE	crm_nouns_lemma	option
OFFRE	crm_nouns_lemma_manual	abonnement
OFFRE#PRODUIT	crm_nouns_lemma	matériel
OFFRE#PRODUIT	crm_nouns_lemma	produit
OFFRE#PRODUIT	crm_nouns_lemma	stock
OFFRE#SERVICE	crm_nouns_lemma	prestation
PRIX	crm_adj_lemma_manual	ch[éeè]re?s?
PRIX	crm_adj_lemma_manual	gratuit?

PRIX	crm_nouns_lemma	devis
PRIX	crm_nouns_lemma	prix
PRIX	crm_nouns_lemma	tarif
PRIX#AUGMENTATION	crm_nouns_lemma	augmentation
PRIX#REDUCTION	crm_from_regex	geste commercial
PRIX#REDUCTION	crm_nouns_lemma	promo
PRIX#REDUCTION	crm_nouns_lemma	promotion
PRIX#REDUCTION	crm_nouns_lemma	réduction
PRIX#REDUCTION	crm_nouns_lemma	remise
PRIX#REDUCTION	crm_nouns_lemma_manual	(re)négociation
PRIX#REDUCTION	crm_verbs_lemma_manual	(re)négocier
SOCIETE#ENTITE#DEPARTEMENT	crm_from_regex	service client
SOCIETE#ENTITE#DEPARTEMENT	crm_from_regex	service technique
SOCIETE#ENTITE#DEPARTEMENT	crm_nouns_lemma	département
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	conseiller
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	dame
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	directeur
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	employé
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	équipe
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	expert
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	interlocuteur
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	opérateur
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	personne
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	personnel
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	pro
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	responsable
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	technicien
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma	vendeur
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma_manual	h[ôo]tesse
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma_manual	téléconseill(er|ère)s?
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma_manual	téléopérat(eur|rice)s?
SOCIETE#ENTITE#INTERLOCUTEUR	crm_nouns_lemma_manual	télévendeu(r|se)s?
SOCIETE#ENTITE#STRUCTURE	crm_nouns_lemma	compagnie
SOCIETE#ENTITE#STRUCTURE	crm_nouns_lemma	enseigne
SOCIETE#ENTITE#STRUCTURE	crm_nouns_lemma	société
SOCIETE#ENTITE#STRUCTURE	crm_nouns_lemma_manual	marque
SOCIETE#LIEU#PHYSIQUE	crm_from_regex	salon/salle d'attente
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	agence
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	boutique
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	centre
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	espace
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	guichet
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	magasin
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	réception
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	salle
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	salon
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	siège
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma	site
SOCIETE#LIEU#PHYSIQUE	crm_nouns_lemma_manual	loca(l|us?x?)
SOCIETE#LIEU#VIRTUEL	crm_from_regex	site internet
TRANSACTION#ACHAT	crm_nouns_lemma	achat
TRANSACTION#ACHAT	crm_nouns_lemma	commande
TRANSACTION#ACHAT	crm_nouns_lemma	livraison
TRANSACTION#FACTURE	crm_nouns_lemma	facturation
TRANSACTION#FACTURE	crm_nouns_lemma	facture
TRANSACTION#PAIEMENT	crm_nouns_lemma	paiement
TRANSACTION#PAIEMENT#MODES DE PAIEMENT	crm_from_regex	carte bancaire
TRANSACTION#PAIEMENT#MODES DE PAIEMENT	crm_from_regex	carte bleue
TRANSACTION#PAIEMENT#MODES DE PAIEMENT	crm_nouns_lemma	cb
TRANSACTION#PAIEMENT#MODES DE PAIEMENT	crm_nouns_lemma	pr[éeè]l[éeè]vements?

TRANSACTION#PAIEMENT#MODES DE PAIEMENT	crm_nouns_lemma	virement	
TRANSACTION#PAIEMENT#MODES DE PAIEMENT	crm_nouns_lemma_manual	cotisation	
TRANSACTION#PAIEMENT#MODES DE PAIEMENT	crm_verbs_lemma_manual	payer	
TRANSACTION#PAIEMENT#MODES DE PAIEMENT	crm_verbs_lemma_manual	verser	
TRANSACTION#REMBOURSEMENT	crm_nouns_lemma	remboursement	

Annexe 3
Règles symboliques

Cette annexe présente le fichier des règles composant la méthode symbolique, classées par catégories. En en-tête se trouve la déclaration des classes d'annotation utilisées.

```
#######################
# GLOBAL DECLARATIONS #
#######################


firsttok = { type: "CLASS",
value:"fr.ho2s.holmes.annotations.HolmesAnnotations$IsFirstTokenAnnotation" }
command = { type: "CLASS", value:
"fr.ho2s.holmes.annotations.SemanticTypeAnnotations$ClassificationCommandAnnotatio
n" }
lasttok = { type: "CLASS", value:
"fr.ho2s.holmes.annotations.HolmesAnnotations$IsLastTokenAnnotation" }
numtype = { type: "CLASS", value:
"fr.celi.hybrid.custom.CustomAnnotations$NumberTypeAnnotation" }
pos = { type: "CLASS", value:
"edu.stanford.nlp.ling.CoreAnnotations$PartOfSpeechAnnotation" }
ta = { type: "CLASS", value: "fr.celi.hybrid.HybridAnnotations$TokenAnnotation" }
sa = { type: "CLASS", value:
"fr.ho2s.holmes.annotations.SemanticTypeAnnotations$SemanticFeatureAnnotation" }
groupString = {type: "CLASS", value:
"fr.ho2s.holmes.annotations.HolmesGroupAnnotations$TokenGroupsStringAnnotation"}
pol = {type: "CLASS", value:
"edu.stanford.nlp.ling.CoreAnnotations$PolarityAnnotation"}

# Define ruleType to be over tokens
ENV.defaults["ruleType"] = "tokens"

# Case insensitive pattern matching (see java.util.regex.Pattern flags)
ENV.defaultStringPatternFlags = 10

#######################
# COMMAND APPLICATION #
#######################

ENV.defaults["stage"] = 1
```

##
##
#1#ACCUEIL
##
##

##
##
#2#ATTITUDE
##
##

##
##
#3#CONFORT
##
##

##
##
#4#ATTENTE
##
##

#j'ai poireauté plus d'une heure
{ pattern: (([{sa:/ATTENDRE.*/}]) /.+/{0,5} [{sa:/DELAI#TEMPS.*/}]),
  result: (HolmesGroup($1,"command","action:BOOST;target_cat:4")),
  name: "CRM4A" }

##
##
#5#ACCESSIBILITE
##
##

```
#notre interlocutrice n'est jamais disponible
{ pattern: ( ([{sa:/SOCIETE.*/}]) /.+/{0,5} [{sa:/DISPONIBILITE.*/} &
{pos:/NC|ADJ|NPP/}] ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:5") ),
  name: "CRM5A" }

########################################
########################################
#6#STABILITE
########################################
########################################

#toujours des interlocuteurs différents
{ pattern: ( ([{sa:/SOCIETE#ENTITE#INTERLOCUTEUR.*/}]) /.+/{0,3}
[{lemma:/diff[éeè]rent?e?s?/}] ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:6") ),
  name: "CRM6A" }


########################################
########################################
#7#ASSISTANCE
########################################
########################################




########################################
########################################
#8#PROACTIVITE
########################################
########################################




########################################
########################################
#9#COMPREHENSION
########################################
########################################




########################################
########################################
#10#COMPETENCE
########################################
```

```
########################################


########################################
########################################
#11#FIDELITE
########################################
########################################

#client de/depuis de nombreuses années
{ pattern: ( [{sa:/CLIENT#EN COURS.*/}] /.+/{0,5} [{lemma:/depuis|de/}] /.+/{0,5}
([{word:/[0-9]+/} | {sa:/DELAI#TEMPS.*/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:11") ),
  name: "CRM11A" }

#depuis de nombreuses années client
{ pattern: ( [{lemma:/depuis/}] /.+/{0,5} ([{word:/[0-9]+/} | {sa:/DELAI#TEMPS.*/}])
[{sa:/CLIENT#EN COURS.*/}]),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:11") ),
  name: "CRM11B" }

#chez AAA depuis 3 ans
{ pattern: ( [{lemma:/chez|[aà]/}] [{sa:/SOCIETE#ENTITE#STRUCTURE#NOM.*/} |
{lemma:/vous/}] /.+/{0,5} [{lemma:/depuis/}] /.+/{0,5} ([{word:/[0-9]+/} |
{sa:/DELAI#TEMPS.*/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:11") ),
  name: "CRM11C" }

#depuis 3 ans chez AAA
{ pattern: ( [{lemma:/depuis/}] /.+/{0,5} ([{word:/[0-9]+/} | {sa:/DELAI#TEMPS.*/}])
[{lemma:/chez|[aà]/}] [{sa:/SOCIETE#ENTITE#STRUCTURE#NOM.*/}] ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:11") ),
  name: "CRM11D" }

#lexique de FIDELITE
{ pattern: ( ([{sa:/FIDELITE.*/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:11") ),
  name: "CRM11E" }

########################################
########################################
#12#DEPART
########################################
########################################

#lexique de DEPART
{ pattern: ( ([{sa:/DEPART.*/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:12") ),
```

```
  name: "CRM12A" }


########################################
########################################
#13#ENGAGEMENT
########################################
########################################

#le conseiller n'a pas tenu ses engagements
{ pattern: ( [{lemma:/tenir/}] /.+/{0,5} ([{sa:/ENGAGEMENT.*/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:13") ),
  name: "CRM13A" }

#des promesses non tenues
{ pattern: ( ([{sa:/ENGAGEMENT.*/}]) /.+/{0,5} [{lemma:/tenir/}]),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:13") ),
  name: "CRM13B" }



########################################
########################################
#14#OFFRE
########################################
########################################




########################################
########################################
#15#TARIFS
########################################
########################################

#lexique de PRIX
{ pattern: ( ([{sa:/PRIX.*/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:15") ),
  name: "CRM15A" }


########################################
########################################
#16#ACHAT
########################################
########################################

#je me suis encore faite avoir
```

```
{ pattern: ( ([{pos:CLO}]) [{lemma:/être/}]? [{pos:ADV}]{0,3} [{lemma:faire}]
[{lemma:avoir}] ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:16") ),
  name: "CRM16A" }



##########################################
##########################################
#17#CONTRAT
##########################################
##########################################




##########################################
##########################################
#18#INFOS
##########################################
##########################################

#vous avez répondu à toutes mes questions
{ pattern: ( [{lemma:/répondre/}] /.+/{0,5} ([{lemma:/question/}])),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:18") ),
  name: "CRM18B" }


##########################################
##########################################
#19#IMAGE
##########################################
##########################################

#a fuir
{ pattern: ( [{word:/à|a/}] ([{lemma:/fuir/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:19") ),
  name: "CRM19A" }


##########################################
##########################################
#20#CONCURRENCE
##########################################
##########################################


#je vais aller voir d'autres marques
{ pattern: ( [{lemma:/autres?/}] /.+/{0,2} ([{sa:/SOCIETE#ENTITE#STRUCTURE/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:20") ),
```

```
  name: "CRM20A" }


#lexique de CONCURRENCE
{ pattern: ( ([{sa:/CONCURRENCE.*/}]) ),
  result: ( HolmesGroup($1,"command","action:BOOST;target_cat:20") ),
  name: "CRM20B" }


########################################
########################################
#21#HORSCATEG
########################################
########################################
```

Annexe 4
Résultats détaillés

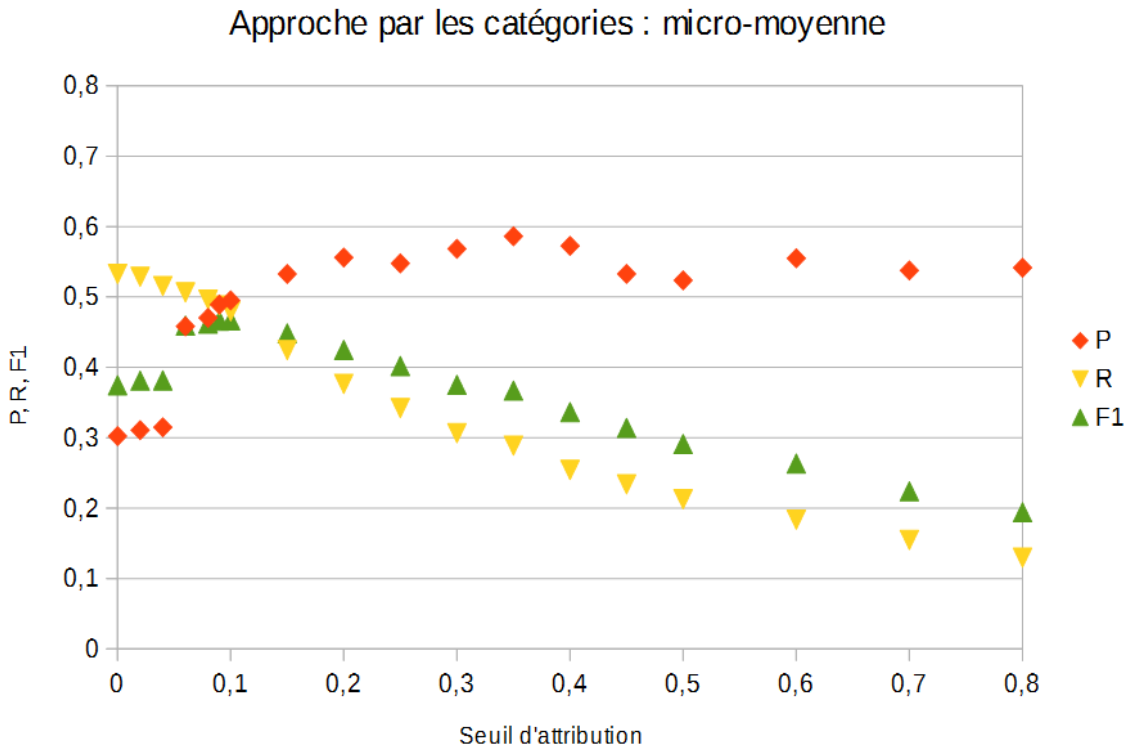
Cette annexe donne les résultats détaillés par configuration, selon trois approches différentes dans la méthode de calcul. Le paramètre portant sur le nombre de catégories maximum attribuées est toujours fixé à 2.

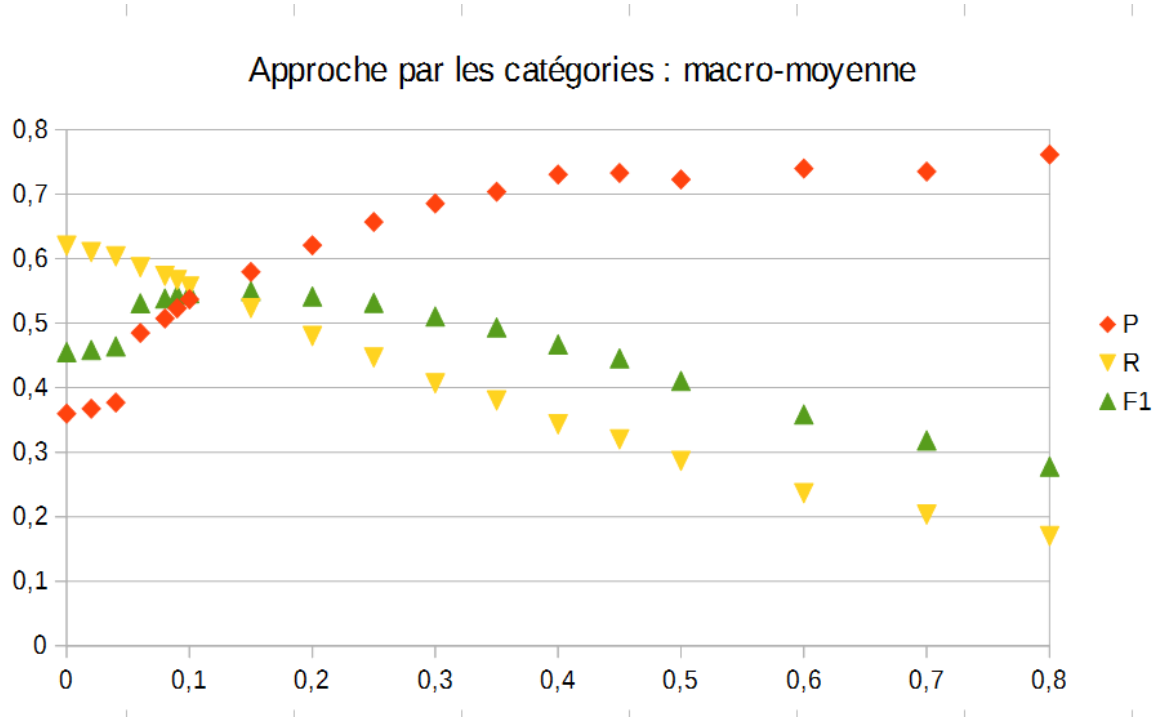
	Approche par les documents					
	Seuil	P	R	F1	A	H
BASIC	0,1	53,55%	53,40%	**51,47%**	45,73%	0,0679
	0,15	54,50%	49,45%	50,25%	45,69%	0,0598
	0,09	51,18%	53,40%	50,13%	43,83%	0,0715
TAXO_1	0,1	54,74%	55,85%	**53,08%**	47,03%	0,0652
	0,15	55,92%	51,70%	51,71%	46,50%	0,0598
	0,09	53,79%	56,56%	**52,84%**	46,48%	0,0670
TAXO_2	0,1	55,45%	56,67%	**53,74%**	47,67%	0,0639
	0,15	56,64%	53,36%	52,76%	47,39%	0,0594
	0,09	54,74%	58,57%	**54,06%**	47,27%	0,0657
SYMBO	0,1	56,16%	57,42%	**54,57%**	48,67%	0,0625
	0,15	57,35%	54,11%	53,59%	48,40%	0,0580
	0,09	55,45%	58,85%	**54,64%**	48,12%	0,0643

		Approche par les catégories												
		Micro-moyenne									Macro-moyenne			
	Seuil	TP	FN	FP	Annotées	Classées	P	R	F1	A	P	R	F1	A
BASIC	0,1	7,19	7,10	7,33	14,29	14,52	49,51%	50,33%	49,92%	33,26%	46,68%	39,22%	39,88%	27,94%
	0,15	6,71	7,57	5,10	14,29	11,81	56,85%	47,00%	51,46%	34,64%	55,80%	35,98%	40,26%	28,69%
	0,09	7,24	7,05	8,05	14,29	15,29	47,35%	50,67%	48,95%	32,41%	44,29%	39,38%	39,26%	27,45%
TAXO_1	0,1	7,62	6,67	7,10	14,29	14,71	51,78%	53,33%	52,55%	35,63%	48,70%	45,13%	44,20%	31,23%
	0,15	7,00	7,29	5,33	14,29	12,33	56,76%	49,00%	52,59%	35,68%	56,09%	40,18%	42,67%	30,03%
	0,09	7,71	6,57	7,57	14,29	15,29	50,47%	54,00%	52,17%	35,29%	47,22%	45,47%	43,84%	30,95%
TAXO_2	0,1	7,71	6,57	6,90	14,29	14,62	52,77%	54,00%	53,38%	36,40%	48,17%	46,25%	45,09%	32,30%
	0,15	7,24	7,05	5,48	14,29	12,71	56,93%	50,67%	53,62%	36,63%	52,75%	40,75%	43,20%	30,81%
	0,09	7,90	6,38	7,48	14,29	15,38	51,39%	55,33%	53,29%	36,32%	47,64%	47,24%	45,13%	32,14%
SYMBO	0,1	7,95	6,33	6,86	14,29	14,81	53,70%	55,67%	54,66%	37,61%	49,52%	47,95%	46,67%	33,51%
	0,15	7,48	6,81	5,43	14,29	12,90	57,93%	52,33%	54,99%	37,92%	53,27%	42,45%	44,84%	32,08%
	0,09	8,10	6,19	7,38	14,29	15,48	52,31%	56,67%	54,40%	37,36%	48,93%	48,81%	46,64%	33,28%

Annexe 5
Seuil d'attribution de catégorie

Nous présentons dans cette annexe l'étude des résultats en fonction du seuil d'attribution de catégorie, afin de fixer celui-ci dans le classifieur selon la meilleure performance.




Approche par les catégories : macro-moyenne
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
P
R
F1

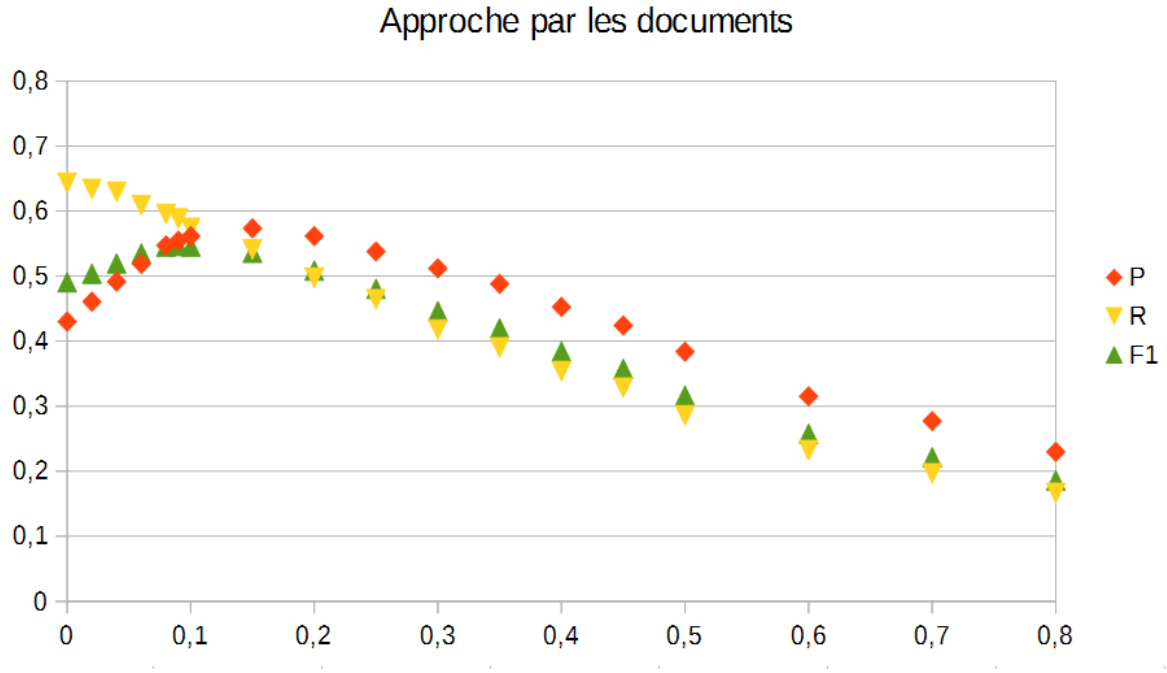

Approche par les documents
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
P
R
F1

Table des matières

MOTS-CLÉS : catégorisation, taxonomie, enrichissement sémantique, méthode hybride, apprentissage automatique, règles symboliques

RÉSUMÉ

Nous avons développé un système de classification multi-catégoriel de textes courts, fondé sur une méthode hybride, afin de déterminer quel est l'apport de l'enrichissement sémantique sur une telle tâche, avec peu de données d'entrainement. Pour cela nous avons tout d'abord constitué différents corpus de documents. Une taxonomie du domaine a été élaborée dans un but d'annotation lexicale sémantique des textes. Par la suite, nous avons développé un système de classification hybride (combinant apprentissage automatique et règles symboliques). Enfin, nous avons mis en place des mesures d'évaluation pour déterminer les performances du classifieur. Nos résultats tendent à montrer que l'enrichissement sémantique est positif car il améliore les performances du classifieur. L'annotation sémantique lexicale permet une meilleure représentation des documents, sur laquelle se basent les méthodes statistique et symbolique. D'autre part, les règles symboliques comblent les lacunes de l'apprentissage automatique.

KEYWORDS : categorization, taxonomy, semantic enrichment, hybrid method, machine learning, symbolic rules

ABSTRACT

We have developed a multi-label classification system for short texts, based on a hybridemethod, in order to explore what impact has the semantic enrichment on such task, with few training data. In this aim, we have first gathered several corpus of documents. A domain taxonomy has been created in order to add semantic lexical annotation to the texts. Then we have developed a hybrid classification system (with machine learning and symbolic rules). Finally, we have defined evaluation measures to determine the classifier's performance. Our results show that semantic enrichment is positive because it improves the classifier's performance. Semantic lexical annotation allows a better document representation, on which the statistical and symbolic methods are based. Furthermore, symbolic rules address the shortcoming of the machine learning.