La désambiguisation des abréviations du domaine médical

Anaïs Koptient

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

La désambiguïsation des abréviations du domaine médical

Anaïs Koptient[1]
(1) CNRS, UMR 8163, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
anais.koptient.etu@univ-lille.fr

RÉSUMÉ

Les abréviations, tout en étant répandues dans la langue, ont une sémantique assez opaque car seulement les premières lettres sont transparentes. Cela peut donc empêcher la compréhension des abréviations, et des textes qui les contiennent, par les locuteurs. De plus, certaines abréviations sont ambiguës en ayant plusieurs sens possibles, ce qui augmente la difficulté de leur compréhension. Nous proposons de travailler avec les abréviations de la langue médicale dans un cadre lié à la simplification automatique de textes. Dans le processus de simplification, il faut en effet choisir la forme étendue des abréviations qui soit correcte pour un contexte donné. Nous proposons de traiter la désambiguïsation d'abréviations comme un problème de catégorisation supervisée. Les descripteurs sont construits à partir des contextes lexical et syntaxique des abréviations. L'entraînement est effectué sur les phrases qui contiennent les formes étendues des abréviations. Le test est effectué sur un corpus construit manuellement, où les bons sens des abréviations ont été définis selon les contextes. Notre approche montre une F-mesure moyenne de 0,888 sur le corpus d'entraînement en validation croisée et 0,773 sur le corpus de test.

ABSTRACT

Disambiguation of abbreviations from the medical domain.

Abbreviations, although commonly used, have quite opaque semantics because only their first letters are transparent. This may prevent from understanding of abbreviations, and of texts they occur within, by speakers. Besides, some abbreviations are ambiguous and have more than one meaning, which increases their understanding difficulty. We propose to work with abbreviations from the medical domain as part of the automatic text simplification. During the simplification process, it is indeed necessary to chose the right expanded form of abbreviations satisfying a given context. We propose to address disambiguation of abbreviations as supervised categorization problem. Descriptors are built from lexical and syntactic contexts of the abbreviations. Training is done on sentences containing expanded forms of the abbreviations. Test is done on corpus built manually, in which the correct senses of abbreviations have been defined according to their contexts. The average F-measure of our approach is 0.888 in cross-validation on the training corpus and 0.773 on the test corpus.

MOTS-CLÉS : Désambiguïsation sémantique, domaine biomédical, abréviations, simplification.

KEYWORDS: Word sense disambiguation, Medical domain, Abbreviations, Simplification.

1 Introduction

Les abréviations sont assez répandues dans les informations et les situations qui nous entourent. Les quelques exemples en (1) illustrent la variété de ces situations.

Volume 3 : Rencontre des Étudiants Chercheurs en Informatique pour le TAL, pages 151–163.
HAL : hal-02786196.

(1) *CAF => Caisse des Allocations Familiales*
PC => Personal Computer
ADP => Aéroports de Paris
TGV => Train à grande vitesse
AVC => Accident Vasculaire Cérébral
IRM => Imagerie par Résonance Magnétique
DP => Dialyse Péritonéale
ACP => analgésie contrôlée par le patient
AC => anhydrase carbonique

Notre compétence linguistique face aux abréviations varie en fonction de la nature de ces abréviations. Ainsi, la signification de certaines abréviations peut nous être connue du fait de leur fréquence d'emploi dans la langue et de leur implication dans le quotidien, comme c'est le cas de *CAF, PC, TGV, ADP* ou même *AVC* et *IRM*. Avec d'autres abréviations, la compréhension est beaucoup moins aisée, comme par exemple avec *DP, ACP* ou *AC*. Dans ces cas, il est nécessaire de disposer de la forme étendue des abréviations pour mieux les comprendre. En effet, les abréviations ont une sémantique très opaque car seulement les premières lettres sont transparentes alors que le reste des mots ne l'est pas. De plus, les abréviations et les mots qui les composent peuvent être spécifiques aux domaines de spécialité, comme le domaine médical dans les exemples que nous utilisons. La sémantique devient alors encore plus opaque. Pour aider le lecteur à bien comprendre la signification des abréviations, et des textes qui les comportent, il est nécessaire de fournir au moins les formes étendues des abréviations. Notons que, dans ce qui suit, nous utilisons de manière équivalente les termes suivants : forme étendue, forme développée, développement et sens des abréviations.

La simplification, automatique ou manuelle, de textes a justement pour objectif de rendre un texte plus lisible et compréhensible. Ainsi, différents guides d'aide à la rédaction de documents simples et accessibles (OCDE, 2015; Ruel J. & L., 2011; UNAPEI, 2019) préconisent, entre autres, de fournir les formes étendues des abréviations. Cela correspond au cadre dans lequel s'inscrit notre travail : simplifier automatiquement les documents en langue française pour les rendre plus faciles à comprendre. Nous nous intéressons ici plus particulièrement à l'explicitation et la simplification d'abréviations. Une autre particularité est que la simplification est effectuée avec les documents techniques du domaine médical. Ce domaine, tout en touchant intimement à notre vie, manipule typiquement de nombreux termes et abréviations spécialisés.

La simplification d'abréviations repose sur la disponibilité et l'exploitation de ressources dédiées, où les abréviations sont associées avec leurs formes étendues, comme dans les exemples en (1). Cependant, de nombreuses abréviations peuvent être ambiguës et avoir plusieurs développements possibles. Par exemple, dans la langue générale, l'abréviation *PC* peut signifier *Personal Computer* mais aussi *Parti Communiste*, et seul le contexte pourra indiquer quel sens et quelle forme étendue correspondante sont corrects pour une occurrence donnée de cette abréviation. La même situation se vérifie dans les langues de spécialité : les abréviations techniques peuvent également être ambiguës et avoir plusieurs développements possibles. C'est notamment le cas des abréviations en (2).

(2) *DP* : *Dialyse Péritonéale*, *Dilatation Pneumatique*, *Dysménorrhée Primaire* ;
ACP : *amplification en chaîne par polymérase*, *analgésie contrôlée par le patient* ;
AC : *ablation par cathéter*, *âge corrigé*, *acétate de cyprotérone*, *anhydrase carbonique*, *anthracycline et cyclophosphamide* ;
ADP : *accès douloureux paroxystiques*, *adénosine diphosphate*.

Le nombre d'abréviations ambiguës est potentiellement élevé et peut donc concerner un nombre assez important de phrases. Le fait qu'une abréviation ait plusieurs développements possibles devient donc problématique, car le système de simplification doit sélectionner la forme développée correcte pour une abréviation étant donné le contexte. Il est donc nécessaire d'effectuer la désambiguïsation pour simplifier une phrase donnée correctement.

Dans ce qui suit, nous présentons d'abord un état de l'art sur la désambiguïsation de mots et termes (section 2). Ensuite, nous décrivons notre démarche pour désambiguïser les abréviations du domaine médical en français (section 3). Les résultats sont présentés et discutés dans la section 4. Nous concluons avec quelques pistes d'amélioration et les perspectives pour les travaux futurs.

2 État de l'art

L'intérêt pour la désambiguïsation sémantique a été manifesté très tôt par la communauté de TAL : dès l'arrivée des premiers programmes informatiques au début des années 50 (Ide & Véronis, 1998). Actuellement, la désambiguïsation est utilisée en pré-traitement de plusieurs autres programmes et applications, comme par exemple :

— la traduction automatique car un mot de la langue source peut avoir plusieurs traductions possibles dans la langue cible, en fonction de sa sémantique (Vickrey *et al.*, 2005; Miháltz, 2005; Li & Li, 2004; Specia, 2005; Lim & Tang, 2004; Marvin & Koehn, 2018; Tang *et al.*, 2018; Parameswarappa & Narayana, 2011; Brown *et al.*, 1991) ;

— l'extraction d'information car, lors de la recherche de mots-clés, il est important de pouvoir éliminer les mots-clés qui n'ont pas le sens recherché (Stokoe *et al.*, 2003; Zhong & Ng, 2012; Whaley, 1999; Krovetz, 2002; Stokoe & Tait, 2002) ;

— l'analyse grammaticale lors de l'annotation en parties du discours, par exemple pour éviter d'annoter de manière erronée un mot dont l'homonyme n'a pas la même partie du discours (Bikel, 2000).

Les travaux que nous mentionnons ici concernent la désambiguïsation effectuée essentiellement sur les données de la langue médicale, y compris sur les abréviations du domaine médical, car c'est dans ce domaine que nous positionnons notre travail.

Pour la présentation de travaux existants, nous différencions les méthodes non supervisées (section 2.1) et les méthodes supervisées (section 2.2). L'avantage de méthodes non supervisées est qu'elles ne demandent pas de gros corpus annotés pour l'entraînement de systèmes et sont donc plus faciles à mettre en place. En revanche, elles obtiennent théoriquement de moins bons résultats que les méthodes supervisées car elles utilisent moins de connaissances et disposent d'une plus petite fraction de la vérité de terrain (Zhou & Han, 2005). Nous verrons cependant que les chercheurs mettent en place différentes approches afin de réduire l'effort nécessaire à la création de données annotées pour l'entraînement, grâce à l'exploitation de corpus parallèles, de corpus faiblement annotés ou encore de connaissances fournies par des ressources existantes, comme l'UMLS (Lindberg *et al.*, 1993).

2.1 Méthodes non supervisées

Dans un travail, les chercheurs utilisent un corpus parallèle anglais-allemand (*Spinger Corpus of Medical Abstracts*) pour effectuer la désambiguïsation de termes médicaux (Widdows *et al.*, 2003). L'avantage de ce type de corpus est qu'un mot ambigu dans une langue ne l'est pas toujours dans

une autre langue : la mise en parallèle de ces deux langues permettrait donc de désambiguïser les occurrences dans l'une des langues. L'annotation est effectuée automatiquement en utilisant les CUI [1] de l'UMLS. Ainsi, pour un terme ambigu dans le résumé en anglais, les auteurs cherchent sa traduction dans le résumé correspondant en allemand. Les traductions sont gardées uniquement si (1) seulement un CUI est assigné à n'importe quel terme du résumé en allemand et (2) au moins un des termes, auquel le CUI est assigné dans le résumé en allemand, n'est pas ambigu. De cette manière, la désambiguïsation est effectuée pour les termes ambigus en anglais et les termes ambigus en allemand, tout en utilisant le même corpus de textes parallèles. Pour les termes en anglais, cette méthode donne une précision de 81 % et un rappel de 18 %. Pour la désambiguïsation des termes en allemand, elle donne une précision de 66 % et un rappel de 22 %.

Une autre méthode proposée par le même groupe de chercheurs (Widdows *et al.*, 2003) repose sur les collocations. En effet, il a été observé que les mots et termes ambigus ont tendance à avoir plusieurs collocations, parmi lesquelles une collocation donnée peut correspondre à un sens donné (Yarowsky, 1993). Pour rendre cette propriété totalement non supervisée, les auteurs utilisent également les CUI de l'UMLS, associés alors aux sens (un CUI est supposé correspondre à un sens), pour chaque mot ambigu afin de déterminer ses collocations. Cette méthode ne fournit pas un très bon rappel (3 % sur le corpus en anglais et 1 % sur le corpus en allemand), mais elle donne une précision assez élevée : 79 % pour les termes en anglais et 82 % pour les termes en allemand.

Enfin, une dernière méthode proposée par ce groupe de chercheurs (Widdows *et al.*, 2003) consiste en l'utilisation de termes qui sont liés par des relations conceptuelles contenues dans les tables MRREL et MRCXT de l'UMLS. Ainsi, pour chaque sens d'un mot ambigu w, la méthode cherche les termes qui sont liés à ce sens (également représenté par un CUI) dans les fichiers MRREL et MRCXT. Ensuite, pour chaque occurrence du terme ambigu w, le contexte est identifié et chaque mot du contexte est recherché dans les fichiers liés à chaque sens de w : si le mot du contexte fait partie du sens en particulier alors le score de ce sens est incrémenté. À la fin du processus, le sens qui a le score le plus haut est considéré comme celui qui correspond au sens du mot ambigu w. Cette méthode montre une précision entre 71 et 74 % pour la désambiguïsation de termes en anglais, et une précision entre 77 et 79 % pour la désambiguïsation de termes en allemand. Concernant le rappel, il est entre 32 et 49 % pour les termes en anglais, et entre 31 et 58 % pour les termes en allemand.

2.2 Méthodes supervisées

Un travail, effectué sur les données de la langue médicale exploite une méthode en deux étapes (Liu & Lussier, 2001). Lors de la première étape, pour un terme w ambigu, les auteurs définissent automatiquement un corpus étiqueté sémantiquement grâce à trois ressources (*UMLS Metathesaurus*, *MEDLINE* (NLM, 2015) et *Clinical Data Repository*). La seconde étape correspond alors à la construction d'un classifieur dédié à la désambiguïsation avec plusieurs algorithmes d'apprentissage supervisé (*Naive Bayes*, *Decision List* et *Exemplar-based*). Il s'agit donc d'une méthode mixte. Cette méthode méthode montre une *accuracy* entre 75 et 99 %.

D'autres chercheurs utilisent également les informations sur les CUI de l'UMLS (McInnes *et al.*, 2007). La méthode proposée consiste à obtenir les CUI des termes qui se trouvent dans la même fenêtre contextuelle que le terme ambigu. Dans ce travail, la fenêtre peut correspondre à la phrase dans laquelle se trouve le mot ambigu ou même au résumé complet. Lorsque les CUI contextuels sont

1. *Concept Unique Identifier* : il s'agit d'un code unique qui représente un concept (ensemble de termes sémantiquement équivalents) défini dans le Metathesaurus de l'UMLS.

définis, ils sont assignés manuellement au terme ambigu. Ensuite, le système calcule la fréquence qui correspond au nombre de fois où ce CUI apparaît dans le même contexte que le terme ambigu. Les chercheurs utilisent l'algorithme d'apprentissage Naive Bayes en validation croisée, tel qu'implémenté dans la plateforme WEKA (Witten & Frank, 2005).

Une autre approche proposée (Mihăltz, 2005) détermine d'abord, pour chaque terme ambigu, les informations syntaxiques d'autres mots se trouvant dans la même phrase que ce terme ambigu. Ensuite, le système détermine le domaine sémantique, ou le sujet, du paragraphe dans lequel se trouve le terme ambigu, ce qui permet de définir le sens de ce terme. Ce travail, effectué sur les données en hongrois et en anglais, montre une précision de 76,39 % pour l'anglais et de 84,2 % pour le hongrois.

Un autre chercheur (Pedersen, 2001) assigne un sens à un terme ambigu en exploitant les bigrammes qui se trouvent dans le contexte de ce terme. Différents algorithmes de WEKA sont alors utilisés (les arbre de décision *J48*, *Decision Stump* et *Naive Bayes*).

Une autre méthode exploite le *bilingual bootstrapping* (Li & Li, 2004). Cette méthode consiste à utiliser un petit volume de données annotées et un grand volume de données non annotées dans deux langues (langue source et langue cible). La méthode construit des classifieurs dans ces deux langues en parallèle et renforce la performance des classifieurs, d'une part en classifiant les données de chaque langue et, d'autre part, en échangeant des informations sur les données classifiées dans les deux langues.

Enfin, dans un dernier travail que nous voulons présenter, les chercheurs travaillent spécifiquement sur la désambiguïsation des abréviations du domaine médical (Stevenson *et al.*, 2009). Ils utilisent plusieurs algorithmes d'apprentissage supervisé (*Vector Space Model*, *Naive Bayes* et *Support Vector Machines*). Plusieurs descripteurs et paramètres sont pris en compte pour créer les modèles de désambiguïsation :

— les collocations avec les bi-grammes et tri-grammes (de lemmes, de formes et de parties du discours), de même que les couples forme/lemme se trouvant dans la même phrase que l'abréviation ambiguë,
— les CUI, selon l'approche de (McInnes *et al.*, 2007),
— l'utilisation de termes de Medical Subject Headings (MeSH), qui sont exploités pour indexer les documents médicaux dans MEDLINE. Ces termes sont associés manuellement aux différents résumés. Ainsi, sont utilisés comme descripteurs les termes MeSH qui sont associés aux résumés dans lesquels se trouvent les termes ambigus.

Cette méthode montre une performance entre 0,954 et 0,990.

3 Méthodologie

Notre méthode est une méthode supervisée qui s'appuie sur cinq classifieurs tels qu'implémentés dans la librairie ScikitLearn (Pedregosa *et al.*, 2011). Nous décrivons les données exploitées, les paramètres de la méthode et les principes de l'évaluation.

Les abréviations, le corpus d'entraînement et le corpus de test sont issus du corpus CLEAR (Grabar & Cardon, 2018). Il s'agit d'un corpus composé de trois sous-corpus de textes comparables : un sous-corpus de notices de médicaments, un sous-corpus de résumés de revues systématiques et un sous-corpus d'articles d'encyclopédies en ligne gratuites. Ce corpus contient en effet de nombreuses abréviations du domaine médical. Nous avons également constaté que plusieurs de ces abréviations

sont ambiguës, comme les exemples en (2) présentés dans la section 1.

3.1 Ensemble d'abréviations pour la désambiguïsation

Parmi les 1 638 abréviations détectées dans ce corpus, 138 sont ambiguës. Nous travaillons donc avec ces 138 abréviations. Chaque abréviation est associée avec l'ensemble de ses développements connus.

Sur les 138 abréviations ambiguës, 34 ne sont pas exploitables parce que le nombre d'occurrences est trop faible :

— 11 abréviations n'ont qu'une seule occurrence d'un des développements possibles,
— 7 abréviations n'ont que deux occurrences d'un des développements possibles,
— 16 abréviations n'ont que 3 à 5 occurrences d'un des développements possibles.

Les expériences sont donc effectuées avec 104 abréviations.

Les abréviations ont des niveaux d'ambiguïté différents avec 2 à 7 sens possibles. Le tableau 1 indique le nombre de développements possibles pour ces 104 abréviations. Nous pouvons voir que les abréviations avec 2 sens sont les plus fréquentes mais qu'il n'est pas rare d'avoir des abréviations avec plus de 2 sens.

TABLE 1: Ambiguïté des abréviations : nombre de sens ou de développements possibles.

	Nombre	Exemple
2 développements	72	*VC (ventilation conventionnelle, volume courant)*
3 développements	18	*TRC (taux rémission complète, taux réponse complète, temps recoloration cutanee)*
4 développements	10	*TSA (traitement suppression androgénique, traumatisme sonore aigu, travailleur santé autochtone, Trouble Spectre Autisme)*
>4 développements	4	*RC (réadaptation cardiaque, régime cétogène, rémunération conditionnelle, reponse conditionnelle, rythme cardiaque, rapport des cotes, rémission complète)*

3.2 Données de référence

Les documents du corpus CLEAR sont annotés par l'étiqueteur et l'analyseur syntaxique Cordial (Laurent *et al.*, 2009). Les phrases qui comportent les abréviations ambiguës sont exploitées. Nous avons construit deux corpus de référence :

— *Corpus d'entraînement*. Le corpus d'entraînement contient les phrases avec les formes étendues des abréviations ambiguës. Au total, le corpus contient 174 099 phrases. Ces données de référence sont créées automatiquement car les formes étendues des abréviations ne sont pas ambiguës. Cela représente une moyenne de 1 674 phrases par abréviation. Le minimum d'exemples (n=11) est observé avec l'abréviation *TRA (techniques de reproduction assistée, traitement restaurateur atraumatique, traitement de restauration atraumatique)*. Le maximum d'exemples (n=25 885) est observé avec l'abréviation *PA (phosphatase alkaline, pression*

artérielle, Pseudomonas aeruginosa). Concernant les sens des abréviations, nous avons une moyenne de 662 phrases par sens, avec un minimum d'une seule occurrence (19 formes étendues concernées) et un maximum de 25 455 occurrences pour la forme étendue *pression artérielle*. Nous pouvons voir que, selon les abréviations et les sens, le nombre d'exemples disponibles est plus ou moins élevé. Les données d'entraînement ne sont donc pas équilibrées. Les formes étendues des abréviations sont lemmatisées en même temps que les phrases. Ainsi, lorsqu'il existe plus d'une forme flexionnelle possible (comme *groupes hospitaliers* et *groupe hospitalier* pour *GH*) elles sont groupées ensemble. En revanche, les formes dérivationnelles (comme *traitement restaurateur atraumatique* et *traitement de restauration atraumatique* pour *TRA*) ne sont pas groupées ensemble ;

— *Corpus de test*. Le corpus de test comporte les phrases avec les abréviations ambiguës. Au total, le corpus contient 1 665 phrases. 92 des 104 abréviations ambiguës sont présentes dans le corpus de test mais avec des contextes différents. Ces données de référence sont créées manuellement. Pour chaque phrase, la décision sur le bon sens (forme développée) d'une abréviation est prise grâce au contexte de la phrase et, lorsque cela n'est pas suffisant, nous consultons le document duquel est extraite la phrase.

Ainsi, dans le corpus d'entraînement, nous avons les phrases avec les formes étendues (exemple en (3)), alors que, dans le corpus de test, nous avons les phrases avec les abréviations (exemple en (4)).

(3) *Déterminer l'efficacité des interventions comportementales pour traiter une dysménorrhée primaire ou secondaire les unes par rapport aux autres, par rapport à un placebo, à l'absence de traitement ou à des traitements médicaux conventionnels, par exemple les anti-inflammatoires non stéroïdiens (AINS).*

(4) *En raison du caractère limité des résultats soutenant le recours à la neurectomie antéro-sacrée pour la prise en charge de la DP, les risques doivent être rigoureusement mis en balance avec les avantages attendus.*

Pour la création de descripteurs, nous n'utilisons pas de connaissances externes mais uniquement les informations contextuelles contenues dans les phrases du corpus. Ainsi, pour chaque sens de chaque abréviation ambiguë, nous cherchons ses contextes dans une fenêtre de cinq mots à gauche et de cinq mots à droite. Lorsque le sens de l'abréviation est suivi de l'abréviation elle-même entre parenthèses, nous ne prenons pas en compte les parenthèses et l'abréviation se trouvant entre parenthèses dans le contexte de cinq mots à droite, mais passons directement au premier mot se trouvant après la parenthèse. Ainsi, dans la phrase *la différence moyenne entre les groupes était de -0,18 LogMAR (intervalle de confiance (IC) à 95 % statistiquement significatif de -0,32 à -0,04).*, les mots du contexte de droite seront *à*, *95 %*, *statistiquement*, *significatif* et *de*. Les mots lexicaux et les mots grammaticaux sont pris en compte. Le contexte est représenté par les lemmes et par les étiquettes syntaxiques des co-occurrences des abréviations. De plus, la position de chaque co-occurrence est également retenue. Par exemple, pour illustrer cette transformation, nous utilisons la phrase en (3) pour la forme étendue *dysménorrhée primaire* de l'abréviation *DP*. Cette phrase produit les descripteurs suivants pour les contextes gauche et droit : *posi1-gauche_un*, *posi1-gauche_DETIFS*, *posi2-gauche_traiter*, *posi2-gauche_VINF*, *posi3-gauche_pour*, *posi3-gauche_PREP*, *posi4-gauche_comportemental*, *posi4-gauche_ADJFP*, *posi5-gauche_intervention*, *posi5-gauche_NCFP*, *posi1-droite_ou*, *posi1-gauche_COO*, *posi2-droite_secondaire*, *posi2-gauche_ADJSIG*, *posi3-droite_le*, *posi3-gauche_DETDPIG*, *posi4-droite_un*, *posi4-gauche_PIFP*, *posi5-droite_par_rapport_aux*, *posi5-gauche_PREP*. Ensuite, pour l'ensemble de descripteurs, nous calculons si un descripteur donné se trouve dans le contexte concerné

d'une occurrence d'une abréviation donnée. Si c'est le cas, la valeur du descripteur est 1 et si non sa valeur est 0. Le même calcul de descripteurs est effectué sur les phrases du corpus de test.

3.3 Algorithmes pour l'apprentissage supervisé

Nous exploitons les algorithmes implémentés dans la librairie ScikitLearn (Pedregosa *et al.*, 2011) destinée à l'apprentissage supervisé et non supervisé. Nous avons choisi d'utiliser cinq classifieurs pour un apprentissage supervisé :

— SVM Linear et SVM RBF (Platt, 1998). SVM est un algorithme d'apprentissage supervisé qui peut être utilisé à la fois pour la classification et pour la régression. Ici, nous l'utilisons pour la classification. Il s'agit d'un algorithme qui cherche un hyperplan pour mieux séparer les paramètres des classes. Nous utilisons deux noyaux : linéaire et Gaussien (RBF) ;

— Decision Tree (Quinlan, 1993). Un arbre de décision est représenté sous la forme d'un arbre, où une décision possible est située à chaque embranchement. Elle est atteinte ou non en fonction des choix effectués à chaque étape de l'arbre ;

— MultiLayer Perceptron (Rosenblatt, 1961). Un perceptron multicouche est composé de plusieurs couches dans lesquelles circule une information. Les dernières couches représentent la sortie du système ;

— RandomForest (Breiman, 2001). Les forêts d'arbres décisionnels fonctionnent grâce à un apprentissage effectué sur différents arbres de décision entraînés sur des sous-ensembles de données.

L'ensemble de descripteurs (contextes gauche et droit) est exploité avec les algorithmes. Les algorithmes doivent prédire le sens d'une abréviation ambiguë en fonction du contexte où elle apparaît.

3.4 Évaluation

L'entraînement est effectué sur le corpus d'entraînement et le test est effectué avec les 1 665 phrases du corpus de test. Sur le corpus d'entraînement, pour chaque abréviation, nous effectuons une validation croisée à 10 plis. En fonction du nombre de sens, les modèles sont bi-classes ou multi-classes. Les modèles créés sur ce corpus d'entraînement sont appliqués et testés sur le corpus de test. Sur le corpus de test, nous gardons uniquement la première prédiction, dont la probabilité est la plus grande, pour chaque occurrence d'abréviation et la comparons avec les données de référence. Nous calculons les mesures d'évaluation classiques (Sebastiani, 2002) : Précision, Rappel et F-mesure dans leurs versions micro. Nous calculons également la moyenne de ces mesures pour chaque algorithme.

Notre *baseline* correspond à la catégorisation des sens dans la catégorie majoritaire. Les résultats de la baseline sont également évalués en termes de Précision, Rappel et F-mesure.

4 Résultats

Le tableau 2 indique les résultats obtenus avec une validation croisée à 10 plis sur le corpus d'entraînement. Nous pouvons constater que ces résultats sont assez élevés et que l'algorithme Multi-Layer Perceptron (MLP) obtient de meilleurs résultats avec une F-mesure de 0,888. Les valeurs de Précision et de Rappel sont équilibrées pour les différents algorithmes testés. Ces résultats en validation croisée

sont comparables avec les résultats de l'état de l'art (Liu & Lussier, 2001; Mihăltz, 2005). La dernière colonne du tableau indique les performances obtenues avec la *baseline*. Nous voyons que tous les algorithmes exploités montrent des résultats supérieurs à la *baseline*.

TABLE 2: Résultats obtenus sur le corpus l'entraînement par chaque algorithme en validation croisée.

Mesure	SVM Linear	SVM RBF	Decision Tree	MLP	RandomForest	Baseline
Rappel	0,885	0,878	0,897	0,905	0,887	0,822
Précision	0,880	0,849	0,887	0,892	0,871	0,822
F-mesure	0,877	0,856	0,888	0,895	0,873	0,822

Les résultats de désambiguïsation des abréviations dans le corpus de test, obtenus avec différents algorithmes testés, sont présentés dans le tableau 3. Nous voyons que Decision Tree présente les meilleurs résultats avec 0,773 de F-mesure et les valeurs de Précision et de Rappel assez équivalentes. Les autres algorithmes montrent un Rappel beaucoup plus bas, ce qui diminue leurs performances globales. Nous remarquons que les résultats de la *baseline* sont supérieurs aux résultats fournis par les algorithmes.

TABLE 3: Résultats de la désambiguïsation dans le corpus de test.

Mesure	SVM Linear	SVM RBF	Decision Tree	MLP	RandomForest	Baseline
Rappel	0,402	0,398	0,788	0,424	0,402	0,822
Précision	0,797	0,755	0,759	0,763	0,728	0,822
F-mesure	0,534	0,524	0,773	0,545	0,518	0,822

Le tableau 4 indique le nombre d'abréviations qui ont été correctement traitées et désambiguïsées par les différents algorithmes. Nous pouvons voir que Decision Tree arrive à traiter correctement le plus grand nombre d'occurrences (547).

TABLE 4: Nombre total d'occurrences des abréviations classifiées correctement dans le corpus de test.

SVM Linear	SVM RBF	Decision Tree	MLP	RandomForest
441	516	547	523	492

Parmi les abréviations qui sont correctement désambiguïsées avec 100 % de prédictions correctes, nous avons par exemple *DIU (diplôme inter universitaire, dispositif intra utérin)* et *GH (groupe hospitalier, growth hormone)*. Nous voyons deux raisons principales à cela : (1) les sens de ces deux abréviations ont des sémantiques très éloignées et donc des contextes très différents et (2) ces abréviations ont de nombreux exemples dans les données d'entraînement. Pour ces deux raisons, leur désambiguïsation est facilitée. 18 autres abréviations sont dans ce cas également. De plus, leur désambiguïsation s'avère aisée pour tous les algorithmes testés. Plusieurs autres abréviations montrent des performances variées selon les algorithmes, dont les valeurs peuvent aller de 0 à 100 %. Finalement, pour plusieurs abréviations (comme *APS, ASA, HE* et *TRA*), nous n'obtenons

malheureusement pas de bons résultats. Il nous semble que la raison principale est que les exemples pour ces abréviations sont insuffisants voire même absents dans le corpus d'entraînement. Il est donc nécessaire de compléter le corpus d'entraînement avec d'autres occurrences de formes étendues.

Globalement, même si les résultats obtenus sur le corpus de test sont inférieurs à ceux obtenus sur le corpus d'entraînement, ils restent comparables avec les résultats obtenus par d'autres chercheurs dans les travaux existants. Nous pensons que le rappel baisse autant entre la validation croisée dans le corpus d'entraînement et le corpus de test parce que les contextes dans lesquels se trouvent les abréviation et leurs versions étendues sont différents. Ces contextes permettent cependant d'effectuer la désambiguïsation de manière très efficace. Étant donné les résultats obtenus, l'algorithme Decision Tree semble être le plus approprié pour effectuer la tâche de désambiguïsation. Un de ses points forts est de garder les valeurs équilibrées pour la Précision et le Rappel. Nos résultats indiquent cependant qu'il est nécessaire d'apporter plusieurs améliorations à ce travail. L'amélioration la plus importante consiste à enrichir le corpus d'entraînement avec d'autres exemples pour les abréviations et les sens qui n'en disposent pas suffisamment actuellement.

5 Conclusion et discussion

Nous avons présenté notre travail sur la désambiguïsation d'abréviations du domaine médical. Après étude des différents moyens pour effectuer la désambiguisation, nous avons proposé d'utiliser une approche par catégorisation supervisée. L'exploitation de ressources sémantiques, comme la terminologie MESHS utilisée dans un travail existant (Stevenson *et al.*, 2009), reste une perspective pour les travaux futurs. De plus, nous disposons d'un nombre assez importants d'exemples. Ainsi, l'entraînement est effectué sur des phrases, obtenues à partir du corpus CLEAR, qui contiennent les formes étendues, et donc non ambiguës, des abréviations. Le test est effectué sur un corpus construit manuellement, où les bons sens des abréviations ont été définis selon leurs contextes phrastiques ou, si nécessaire, la consultation du document d'origine. L'utilisation de ces deux types d'évaluation (validation croisée et corpus de test) montre que, bien que les résultats en validation croisée soient prometteurs, ce n'est pas forcément le cas lorsque le corpus de test est utilisé. L'approche d'apprentissage supervisé (algorithme Decision Tree) montre actuellement une F-mesure moyenne de 0,888 sur le corpus d'entraînement en validation croisée et 0,773 sur le corpus de test. Les résultats montrés par la *baseline*, où l'on assigne les sens à la catégorie majoritaire, sont supérieurs dans le corpus de test. Nous pensons que le déséquilibre entre les catégories devrait être réduit pour disposer de plus d'exemples et obtenir de meilleurs résultats.

Nous avons ainsi plusieurs pistes pour le travail à venir :

— compléter le corpus d'entraînement avec d'autres exemples, notamment tirés de documents médicaux autre que le corpus CLEAR, ce qui permettraient d'améliorer les résultats pour plusieurs abréviations actuellement sous-représentées,
— compléter l'ensemble d'abréviations avec d'autres développements possibles, ce qui peut conduire à l'augmentation de l'ambiguïté de certaines abréviations (plus de sens connus) et à l'augmentation du nombre d'abréviations ambiguës,
— mettre à jour les modèles et créer de nouveaux modèles pour la désambiguïsation d'abréviations.

Finalement, ce système de désambiguïsation sera intégré dans un système plus global dédié à la simplification de textes techniques du domaine médical.

Remerciements

La présente publication s'inscrit dans le projet *CLEAR (Communication, Literacy, Education, Accessibility, Readability)* financé par l'ANR sous la référence ANR-17-CE19-0016-01. Nous remercions les relecteurs pour leurs remarques constructives.

Références

BIKEL D. M. (2000). A statistical model for parsing and word-sense disambiguation. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora : Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, p. 155–163, USA : Association for Computational Linguistics. DOI : 10.3115/1117794.1117814.

BREIMAN L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

BROWN P. F., DELLA PIETRA S. A., DELLA PIETRA V. J. & MERCER R. L. (1991). A statistical approach to sense disambiguation in machine translation. In *Speech and Natural Language : Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*.

GRABAR N. & CARDON R. (2018). Clear – simple corpus for medical French. In *Workshop on Automatic Text Adaption (ATA)*, p. 1–11.

IDE N. & VÉRONIS J. (1998). Introduction to the special issue on word sense disambiguation : The state of the art. *Computational Linguistics*, **24**(1), 1–40.

KROVETZ R. (2002). On the importance of word sense disambiguation for information retrieval.

LAURENT D., NÈGRE S. & SÉGUÉLA P. (2009). L'analyseur syntaxique Cordial dans Passage. In *Traitement Automatique des Langues Naturelles (TALN)*.

LI H. & LI C. (2004). Word translation disambiguation using bilingual bootstrapping. *Computational Linguistics*, **30**(1), 1–22. DOI : 10.1162/089120104773633367.

LIM L. T. & TANG E. K. (2004). Building an ontology-based multilingual lexicon for word sense disambiguation in machine translation.

LINDBERG D., HUMPHREYS B. & MCCRAY A. (1993). The Unified Medical Language System. *Methods Inf Med*, **32**(4), 281–291.

LIU H. & LUSSIER Y. (2001). Disambiguating ambiguous biomedical terms in biomedical narrative text : An unsupervised method. *Journal of Biomedical Informatics*, **34**, 249–261. DOI : 10.1006/jbin.2001.1023.

MARVIN R. & KOEHN P. (2018). Exploring word sense disambiguation abilities of neural machine translation systems (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1 : Research Papers)*, p. 125–131, Boston, MA : Association for Machine Translation in the Americas.

MCINNES B. T., PEDERSEN T. & CARLIS J. (2007). Using umls concept unique identifiers (cuis) for word sense disambiguation in the biomedical domain. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, **2007**, 533–537.

MIHÁLTZ M. (2005). Towards a hybrid approach to word-sense disambiguation in machine translation.

NLM (2015). *Medline : medical literature on-line*. National Library of Medicine, Bethesda, Maryland. www.ncbi.nlm.nih.gov/sites/entrez.

OCDE (2015). *Guide de style de l'OCDE Troisième édition : Troisième édition*. OECD Publishing.

PARAMESWARAPPA S. & NARAYANA V. (2011). Article : Kannada word sense disambiguation for machine translation. *International Journal of Computer Applications*, **34**(10), 1–8. Full text available.

PEDERSEN T. (2001). A decision tree of bigrams is an accurate predictor of word sense. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PLATT J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning* : MIT Press.

QUINLAN J. (1993). *C4.5 Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.

ROSENBLATT F. (1961). *Principles of Neurodynamics : Perceptrons and the Theory of Brain Mechanisms*. Washington DC : Spartan Books.

RUEL J., KASSI B. M. A. C. & L. M.-M. S. (2011). *Guide de rédaction pour une information accessible*. Gatineau : Pavillon du Parc.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

SPECIA L. (2005). A hybrid model for word sense disambiguation in english-portuguese machine translation. In *IN PROCEEDINGS OF THE 8TH RESEARCH COLLOQUIUM OF THE UK SPECIAL-INTEREST GROUP IN COMPUTATIONAL LINGUISTICS*, p. 71–78.

STEVENSON M., GUO Y., ALAMRI A. & GAIZAUSKAS R. (2009). Disambiguation of biomedical abbreviations. In *Proceedings of the BioNLP 2009 Workshop*, p. 71–79, Boulder, Colorado : Association for Computational Linguistics.

STOKOE C., OAKES M. P. & TAIT J. (2003). Word sense disambiguation in information retrieval revisited. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, p. 159–166, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/860435.860466.

STOKOE C. & TAIT J. (2002). Trec 2002 web track "automated word sense disambiguation for internet information retrieval".

TANG G., SENNRICH R. & NIVRE J. (2018). An analysis of attention mechanisms : The case of word sense disambiguation in neural machine translation. *CoRR*, **abs/1810.07595**.

UNAPEI (2019). *L'information pour tous*. UNAPEI.

VICKREY D., BIEWALD L., TEYSSIER M. & KOLLER D. (2005). Word-sense disambiguation for machine translation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, p. 771–778, USA : Association for Computational Linguistics. DOI : 10.3115/1220575.1220672.

WHALEY J. M. (1999). An application of word sense disambiguation to information retrieval.

WIDDOWS D., PETERS S., CEDERBERG S., CHAN C.-K., STEFFEN D. & BUITELAAR P. (2003). Unsupervised monolingual and bilingual word-sense disambiguation of medical documents using umls. In *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine - Volume 13*, BioMed '03, p. 9–16, USA : Association for Computational Linguistics. DOI : 10.3115/1118958.1118960.

WITTEN I. & FRANK E. (2005). *Data mining : Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

YAROWSKY D. (1993). One sense per collocation. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, p. 266–271, USA : Association for Computational Linguistics. DOI : 10.3115/1075671.1075731.

ZHONG Z. & NG H. T. (2012). Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics : Long Papers - Volume 1*, ACL '12, p. 273–282, USA : Association for Computational Linguistics.

ZHOU X. & HAN H. (2005). Survey of word sense disambiguation approaches.