Résolution des coréférences dans des comptes rendus cliniques
Une expérimentation issue du défi i2b2/VA 2011

Pierre Zweigenbaum[1] Guillaume Wisniewski[1,2] Marco Dinarelli[1]
Cyril Grouin[1] Sophie Rosset[1]

[1] LIMSI–CNRS [2] Université Paris Sud

BP 133 – 91403 Orsay, France
{pierre.zweigenbaum, guillaume.wisniewski, marco.dinarelli, cyril.grouin, sophie.rosset}@limsi.fr

Résumé

Nous présentons les expérimentations réalisées en matière de résolution des coréférences médicales lors de notre participation au défi international i2b2/VA 2011. Nous avons utilisé à la fois des approches à base de règles et par apprentissage. Nous avons d'abord testé un modèle par apprentissage qui détecte les singletons, concepts qui n'occurrent dans aucune chaîne de coréférence. Nous avons ensuite établi trois méthodes à base de règles pour détecter les paires de coréférences. Les méthodes à base de règles ont permis l'obtention des meilleurs résultats sur les corpus d'entraînement comme sur les corpus de l'évaluation officielle, selon une F-mesure moyenne calculée via B^3, MUC, BLANC et CEAF. Notre meilleur système s'est ainsi classé 4^e sur 8 participants sur le corpus ODIE (F = 0,802) et 9^e sur 20 participants sur le corpus i2b2/VA (F = 0,856).

Mots Clef

Détection de coréférences, anaphore, antécédent.

Abstract

We present the experiments we ran on medical coreference resolution while participating in the international i2b2/VA 2011 challenge. We used both rule-based and machine-learning based approaches. We first tested a machine-learning based model which detects singleton mentions, i.e., concept mentions that do not occur in any coreference chain. We then designed three rule-based methods, to detect co-referring mention pairs. Rule-based methods obtained the best results on both training and official evaluation corpora, according to an average F-measure based on B^3, MUC, BLANC, and CEAF. Our best system ranked 4th out of 8 participants on the ODIE corpus (F = 0.802) and 9th out of 20 participants on the i2b2/VA corpus (F = 0.856).

Keywords

Coreference detection, anaphora, antecedent.

1 Introduction

La résolution des coréférences est un problème bien connu en extraction d'information [1, 2, 3, 4, 5]. L'objectif consiste à déterminer si deux expressions (ou « mentions ») dans un texte réfèrent ou pas à la même entité du domaine de discours (un exemple est donné à la page suivante dans le tableau 1). Cela permet par exemple de connaître le nombre de personnes différentes mentionnées dans un texte. La résolution des coréférences en domaine médical est une nouvelle tâche exploratoire avec de nombreuses applications possibles, dont la principale est l'extraction d'information ([6] fournit une revue de l'état de l'art). Alors que l'étude des documents cliniques concerne généralement la détection des concepts médicaux ou encore des prescriptions médicamenteuses (objets des éditions 2009 et 2010 du défi i2b2 [7, 8]), traiter la résolution des coréférences de manière à fournir des chaînes de coréférence devrait permettre aux médecins d'accéder au contenu des documents cliniques de manière plus précise.
Dans cet article, nous présentons les expérimentations que nous avons effectuées en matière de résolution des coréférences dans le domaine médical, lors de notre participation à l'édition 2011 du défi i2b2/VA [9]. Après un bref état de l'art (section 2) et une présentation du défi et de ses corpus (section 3), nous examinons la problématique de résolution des coréférences et son instanciation dans le cadre de ces corpus (section 4). Nous décrivons ensuite les méthodes que nous avons mises au point (section 5) puis nous détaillons l'évaluation que nous avons réalisée (section 6), discutons ses résultats et concluons (section 7).

2 État de l'art

Résolution des coréférences. Comme dans beaucoup de tâches en traitement automatique des langues, différentes approches ont été proposées pour résoudre les coréférences. Les premières approches proposées étaient fondées sur des heuristiques, principalement implémentées sous la forme de règles et d'expressions régulières. Ces approches

reposaient sur des contraintes linguistiques et nécessitaient souvent une analyse linguistique [1, 2], où les contraintes principales sont fondées sur une compatibilité syntaxique et sémantique (genre, humain *vs.* non-humain, etc.).
Depuis les années 1990, la plupart des méthodes proposées reposent sur des approches par apprentissage supervisé. Nous pouvons classer les approches proposées en deux grands groupes : (i) le modèle paire de mentions et (ii) le modèle regroupement et classement. Ces deux approches sont similaires et reposent sur une tâche de classification. Dans le modèle « paire de mentions » [10, 11], le classifieur décide, pour chaque candidat antécédent, si la mention courante renvoie à cet antécédent. Ceci conduit à sélectionner un ensemble de paires de mentions qui sont ensuite regroupées. Cette approche souffre cependant de deux problèmes : (i) le déséquilibre entre exemples d'entraînement positifs et négatifs et (ii) la classification de chaque paire de mentions est réalisée indépendamment des autres. Le modèle « regroupement et classement » a été proposé afin de pallier partiellement le problème d'indépendance des mentions. Dans ce modèle, le classifieur doit décider si une mention est coréférente avec une chaîne en cours de construction [3]. Les caractéristiques généralement utilisées peuvent être divisées en quatre catégories principales : appariement de chaînes, puis niveaux grammatical, sémantique et positionnel.
Plus récemment, des approches plus simples ont été proposées, prouvant que des méthodes déterministes pouvaient également produire de bons résultats [4]. D'après ces travaux, Raghunathan *et al.* [5] ont proposé une approche dite « par crible » uniquement fondée sur des modèles déterministes, permettant l'obtention de résultats comparables si ce n'est meilleurs que ceux des autres systèmes.

Résolution des coréférences dans le domaine médical. Comme souligné par Zheng *et al.* [6], quelques travaux seulement ont porté sur la résolution des coréférences en domaine médical. Nous estimons que le manque de données annotées et partagées a joué un rôle dans cette situation. C'est pour cette raison que nous avons inscrit notre travail dans le cadre d'une participation à une campagne d'évaluation. He [12] a travaillé sur la résolution de coréférences dans des comptes rendus d'hospitalisation, proposant une approche fondée sur des techniques de classification par apprentissage. Les caractéristiques utilisées sont assez standards hormis celles extraites du Metathesaurus UMLS [13].

3 Le défi i2b2/VA 2011 et ses corpus

3.1 Le défi i2b2/VA 2011

La première tâche du défi i2b2/VA 2011 consistait à résoudre les coréférences entre expressions du même type sémantique dans des corpus de comptes rendus cliniques. Deux corpus ont été fournis : un premier corpus dédié au défi i2b2 et un second provenant du projet ODIE [14]. Ces corpus étaient pré-annotés en concepts médicaux, qui constituaient les « mentions » entre lesquelles les coréférences étaient à rechercher. Chaque corpus utilise cependant son propre jeu de types sémantiques : cinq catégories de concepts dans le corpus i2b2 et dix catégories dans le corpus ODIE. Les sorties attendues consistent en des chaînes de coréférence, avec les informations de position dans le document et le type sémantique commun à la chaîne de coréférence.

1. The patient$_1$ is a 67 year old female$_1$ who was transferred for cardiac catheterization$_2$...
2. She$_1$ underwent cardiac catheterization$_2$ today which revealed...
3. An exercise tolerance test with Thallium$_3$ was to be performed in order to ...
4. On 24-1-2012 , she$_1$ underwent a Persantine Thallium study$_3$...

TAB. 1 – Exemples de coréférences. Les *mentions* sont soulignées. La coindexation de plusieurs mentions indique qu'elles font référence à la même *entité*, et donc qu'elles appartiennent à la même *chaîne de coréférence*.

3.2 Description des corpus

Ayant effectué nos travaux d'identification de coréférences dans le cadre d'une campagne d'évaluation, la diversité des types de documents, de leurs origines et de leurs modalités d'annotation dans les corpus d'entrée ont constitué les principales difficultés auxquelles nous avons dû faire face.

Le corpus i2b2. Ce corpus se compose de 292 comptes rendus médicaux — documents de sortie et notes de suivi opératoire — provenant de plusieurs départements médicaux d'hôpitaux nord-américains. Des annotations de référence pour les concepts et les chaînes de coréférences ont été fournies dans le corpus d'entraînement selon cinq catégories de concepts : *person, pronoun, problem, test* et *treatment*. Les trois derniers types correspondent à ceux traités dans l'édition 2010 du défi [15]. Nous avons segmenté selon une répartition 66/33 % le corpus d'entraînement i2b2 en sous-corpus de développement et de test pour effectuer nos expériences et régler nos systèmes.

Le corpus ODIE. Ce corpus se compose de plusieurs types de documents provenant de deux institutions (Mayo Clinic et University of Pittsburgh Medical Center). Le corpus intègre un total de 97 documents, qui relèvent de plusieurs types [1]. Nous avons considéré ces différents corpus comme étant distincts les uns des autres du fait de leurs différences trop importantes. Des annotations de référence étaient fournies pour les concepts et les chaînes de coréférence dans le corpus d'entraînement selon dix catégories de concepts : *anatomicalsite, diseaseorsyndrome, indicatorreagentdiagnosticaid, laboratoryortestresult, none, or-*

1. Comptes rendus médicaux *(Clinical)*, notes de pathologie *(Pathology)*, documents de sortie *(Discharge)*, autres documents *(Other)*, notes de radiologie *(Radiology)*, et documents de chirurgie *(Surgery)*.

ganortissuefunction, other, people, procedure, signorsymptom. Nous avons segmenté ces six corpus d'entraînement en développement et test, sur la base d'une proportion de 80/20 % en raison du faible nombre de documents disponibles.

4 Analyse linguistique et en corpus des coréférences

4.1 Analyse linguistique des coréférences

La tâche de résolution des coréférences prend en entrée des expressions appelées « mentions » telles que *the surgical intervention* ou *she*. Une expression référente renvoie à une « entité » du discours (par exemple, une occurrence d'une revascularisation, ou une personne particulière devant être traitée à l'hôpital). Les expressions qui renvoient à la même entité sont dites coréférentes. La résolution de coréférences s'appuie sur de multiple types de connaissances linguistiques et du domaine. D'un point de vue linguistique, les mentions peuvent être réparties selon le type d'expression qu'elles utilisent : (i) les syntagmes nominaux (*the antibiotics*) ; (ii) Les mots grammaticaux qui agissent comme substituts des syntagmes nominaux : pronoms personnels (*he*, *she*, *us*...), pronoms possessifs (*his*, *her*, *our*...), pronoms démonstratifs (*this*, *that*, *these*...), pronoms relatifs (*which*, *who*, *that*...) ; (iii) les noms propres (*Dr. Brian*). Il importe également de noter que certaines expressions ne sont pas référentes, comme le pronom explétif (*it*, *there*). Le rôle du contexte dans la résolution des anaphores souligne également l'importance de la connaissance du domaine dans le processus. L'accord en genre et en nombre (*she*, *her*, *Mrs.*) ou les restrictions de sélection (*the patient was admitted*, *she was treated*) reflètent généralement les propriétés des entités référentes.

Du point de vue du domaine, les expressions coréférentes peuvent être distinguées selon le type d'entités auquel elles réfèrent. Dans les corpus i2b2 et ODIE, des jeux de types sémantiques sont prédéfinis. Les sens des mots sont souvent organisés en une hiérarchie de types (par ex., WordNet). Puisque le sens des mots est habituellement un élément clé pour référer à une entité, la coréférence entre expressions implique généralement l'égalité de leurs types sémantiques (à l'exception des pronoms qui prennent le type sémantique de leur antécédent). De nombreux dispositifs sémantiques et pragmatiques permettent cependant une variation en granularité (du spécifique au générique) ou bien en relation (métonymie, « anaphore pragmatique ») entre antécédent et anaphore. Si nous n'avons pas identifié d'instances de cette dernière dans le corpus d'entraînement, des différences de granularité ont souvent été trouvées : par exemple, *thrombectomy/the procedure* (« thrombectomie » / « la procédure »), ou plus simplement *coronary artery bypass/artery bypass* (« pontage des artères coronaires » / « pontage artériel »).

4.2 Analyse des coréférences en corpus

Avant de commencer le développement de notre système de résolution des coréférences, nous avons effectué une analyse du corpus [2]. Nos principales observations sont résumées dans le tableau 2. Pour des raisons d'espace, nous ne présentons que les statistiques d'un seul corpus, Beth-Partners (issu du corpus i2b2), mais des observations similaires ont été effectuées sur les autres corpus. Notre analyse montre que la structure des chaînes de coréférence et des documents est relativement simple : contrairement aux documents du domaine général dans lesquels des interactions complexes entre plusieurs mentions apparaissent, nous avons remarqué que les documents des corpus étudiés contiennent assez peu de coréférences : la plupart des mentions qui apparaissent dans un document sont des singletons [3] (par exemple, 52 % des mentions du corpus Beth-Partners sont des singletons). De plus, les chaînes de coréférence contiennent peu de mentions (en moyenne entre 4,6 mentions), sauf celles entre personnes (13,6 mentions).

5 Procédure mise en œuvre

Nous avons abordé la résolution des coréférences au moyen d'une procédure en trois étapes qui prend en entrée un jeu de mentions fournies pour un document : (i) filtrer les singletons ; (ii) étant donné le produit cartésien des mentions candidates restantes, déterminer si une paire de mentions coréfère ou pas ; (iii) étant donné les paires de mentions candidates restantes, produire les chaînes de coréférence résultantes. Nous avons testé une méthode par apprentissage pour l'étape (i), et plusieurs méthodes alternatives, par apprentissage ou par règles, pour l'étape (ii). Pour l'étape (iii), nous avons simplement appliqué des règles de transitivité et n'avons pas effectué de vérifications de compatibilité supplémentaires [4]. Pour des raisons de temps, nous n'avons pas testé la combinaison de toutes les méthodes.

5.1 Étape 1 — Détection des singletons

Sur la base de l'analyse présentée au tableau 2, nous avons remarqué qu'en moyenne, la moitié des mentions d'un rapport médical ne renvoie à aucune autre mention, donc qu'il s'agit de singletons. Il devrait être possible de détecter et filtrer préalablement les singletons avant de construire les paires, puis les chaînes de coréférences sur la base des mentions restantes. Suivant cette intuition, nous avons conçu un module de détection des singletons — implémenté comme un modèle binaire fondé sur SVM — qui vise à décider si une mention est un singleton ou pas.

Pour cela, certaines informations contextuelles sont nécessaires, de même que l'information liée à la mention

2. Toutes les statistiques ont été rassemblées sur notre corpus de développement afin de conserver les estimations des taux d'erreur non biaisées lors du test de nos systèmes.

3. Des mentions qui n'entrent dans aucune relation de coréférence.

4. À cette étape, la transitivité peut rassembler des mentions incompatibles.

TAB. 2 – Statistiques sur le corpus Beth-Partners ; σ renvoie à l'écart-type

	tous	person	problem	pronoun	test	treatment
Nb de mentions	26 660	7 110	7 568	1 389	5 086	5 507
Nb de singletons	14 070	729	4 689	857	4 383	3 412
singleton/mention	52,78%	10,25%	61,96%	61,70%	86,18%	61,96%
Nb de chaînes		474	1 043	—	377	835
Nb moy chaînes/doc	16,64 (σ = 13,97)	2,94 (σ = 1,56)	7,29 (σ = 6,31)	—	3,34 (σ = 3,74)	6,23 (σ = 5,79)
Taille moy des chaînes	4,61 (σ = 10,06)	13,57 (σ = 21,84)	2,95 (σ = 1,58)	—	2,30 (σ = .79)	2,65 (σ = 1,38)

elle-même. L'information contextuelle que nous utilisons est liée aux mentions compatibles dans le voisinage de la mention cible — les mots réalisant de telles mentions de même que les mots entre ces mentions — alors que l'information sur la mention cible elle-même est faite des mots réalisant cette mention. Dans tous les cas, les mots sont enrichis d'informations sémantiques extraites à partir des grammaires de l'outil WMatch [16, 17]. Toutes ces informations sont ensuite encodées pour SVM comme un vecteur de caractéristiques. Nous avons notamment utilisé les informations suivantes :

1. Les deux mentions C_1 et C_2 qui précèdent la mention cible avec les mots qui les réalisent W_{C1} et W_{C2} ;
2. La mention cible T avec les mots qui la réalisent W_T ;
3. Les deux mentions C_3 et C_4 qui suivent la mention cible avec les mots qui les réalisent W_{C3} et W_{C4} ;
4. Les mots entre chaque paire de mention, sans se soucier s'ils appartiennent à d'autres mentions : W_{Bi}, $i \in [1 \dots 4]$;
5. Les mots à gauche de C_1 depuis le début de la phrase : LC ;
6. Les mots à droite de C_4 jusqu'à la fin de la phrase : RC ;
7. Le type d'anaphore ML_i, $i \in [1, ..., 4, T]$ des mentions C_i, $i \in [1, ..., 4]$ et de la mention cible T ;
8. La tête H_i, $i \in [1, ..., 4, T]$ du syntagme de chacune de ces mentions.

Le « type d'anaphore » renvoie au fait qu'une mention peut être instanciée explicitement par des mots pleins (anaphore nominale) ou par un pronom (anaphore pronominale). Pour les têtes de syntagme, nous avons considéré tous les mots après avoir supprimé la ponctuation et les mots issus d'un anti-dictionnaire. Ces informations sont similaires à la tâche d'extraction de relation [18] et encodées en vecteur de caractéristiques de la même manière. Le vecteur de caractéristiques correspondant à ces informations est considéré comme une instance positive dans SVM si la mention cible T est un singleton, négative sinon.

5.2 Étape 2 — Catégorisation des paires de mentions candidates

Étape 2a : mots liés dans le corpus d'entraînement (SD). Une première méthode repose sur des connaissances simples créées à partir des données d'entraînement, recensant toutes les paires de mots que l'on a vus en regard dans une paire de coréférence.

Supposons par exemple une paire de coréférence qui implique deux mentions M_1 et M_2 réalisées par les mots W_{M_1} et W_{M_2}. Après avoir supprimé la ponctuation et les mots de l'anti-dictionnaire de W_{M_1} et W_{M_2}, les mots restants sont considérés comme mots clés pour M_1 et M_2 et sont enregistrés dans la base de connaissances comme étant reliés entre eux. La base de connaissances est peuplée avec l'information issue de l'ensemble du jeu d'entraînement. Afin d'améliorer la robustesse, W_{M_1} et W_{M_2} sont d'abord segmentés en unités élémentaires (« tokens »). La base de connaissances est créée lors de l'étape d'entraînement et est ensuite utilisée à l'étape de classification pour détecter les paires. Deux mentions sont considérées comme coréférentes si elles contiennent des mots faisant l'objet d'une relation dans la base de connaissance.

Étape 2b : trois règles et chaînes de caractères (3R).

Étude de corpus supplémentaire. Le tableau 2 présente le nombre de mentions par chaîne selon le type de chaîne : alors que les chaînes relatives aux personnes sont assez longues (plus de dix mentions en moyenne, avec une forte variation), les chaînes impliquant d'autres types de mentions sont assez courtes (deux à trois mentions en moyenne, avec une faible variation). Cette observation, renforcée par le nombre moyen de chaînes dans un document, nous permet de proposer le modèle suivant pour les coréférences apparaissant dans les comptes rendus cliniques du corpus i2b2/VA : chaque document contient généralement une ou deux chaînes longues du type *person* et plusieurs petites chaînes relatives aux *treatments, tests* et *problem*. Après étude du vocabulaire, nous avons remarqué que les mentions impliquées dans les chaînes de type *person* décrivent soit le patient (généralement exprimé comme *the patient*, parfois *pt*), soit le docteur (mentions avec *dr.*, *m.d.* et un nom propre).

L'étude du vocabulaire des mentions impliquées dans une chaîne de coréférence indique aussi que les mentions coréférentes sont généralement similaires, dans le sens où leur distance de Levenshtein est faible : dans de nombreux cas, les mentions sont exactement les mêmes ou partagent la même tête de syntagme (par ex., « gastric carcinoma » et « this carcinoma », « right lower extremity pain » et « the

pain », etc.). Ainsi, dans le corpus Beth-Partners, il y a 9 865 paires de mentions, dont 3 444 où les mentions sont exactement les mêmes (soit 35 %) et 1 096 où les mentions partagent la même tête de syntagme (soit 11 %)[5]. Il importe de noter que ces statistiques ont été calculées sur des paires de mentions dédoublonnées et sous-estiment en conséquence le nombre d'instances de mentions identiques dans les chaînes de coréférence.

Un crible avec trois règles. Sur la base de ces observations, nous avons implémenté — comme système de base — une approche simple à base de règles. Notre approche revient à appliquer successivement un ensemble de règles déterministes. Elle repose sur l'idée proposée par Raghunathan *et al.* [5] : un crible applique des niveaux de modèles de coréférence déterministes, un à la fois, du plus précis au moins précis ; chaque niveau est construit sur la base de la sortie du niveau précédent. Étant donné la rareté des coréférences et la similarité des mentions coréférentes, on espère qu'un faible nombre de règles sera suffisant pour saisir la plupart des chaînes de coréférence. Par ailleurs, si une règle propose plusieurs mentions candidates comme coréférentes avec une mention donnée, priorité est donnée à la mention la plus proche.

Les trois règles de base que nous avons utilisées dans cette méthode sont les suivantes (nous les décrivons par ordre d'application) : (i) *Who_Ref* : chaque mention *who* (« qui ») est liée à la mention précédente la plus proche du type *person* ; (ii) *Which_Ref* : chaque mention *which* (« qui », « que ») est liée à la mention non pronominale précédente la plus proche de type autre que *person* ; (iii) *String_Equality* : les paires de mentions qui sont exactement identiques sont reliées. Un appariement approximatif a également été testé, mais a engendré de moins bons résultats sur le corpus d'entraînement, il a donc été écarté.

Étape 2c : plus de règles, de chaînes de caractères, et d'information conceptuelle (UM).

Construire une représentation des mentions. Pour pallier la fragilité de l'égalité stricte des chaînes de caractères, nous avons commencé par tester des mesures de similarité entre chaînes de caractères (Levenshtein etc.), y compris en utilisant leurs résultats comme attributs pour entraîner un classifieur supervisé. Cependant, les résultats ont vite plafonné, et plutôt que d'explorer davantage les nombreuses mesures de similarité entre chaînes de caractères, nous sommes passés à une méthode plus linguistique liée aux connaissances du domaine. Son but est de produire un résumé schématique de la signification de la mention : à grands traits, la projection de cette signification sur une hiérarchie de concepts.

La distinction entre humain (people, person) et non-humain (la plupart des catégories i2b2/ODIE : problem, test, treatment, anatomicalpart, etc.) est marquée à la fois linguistiquement (quelques pronoms en anglais font la différence entre humain et non-humain) et dans le domaine des comptes rendus médicaux (les humains sont les acteurs et les patients des événements, alors que les non-humains sont des objets et événements médicaux). Nous avons de ce fait considéré différemment les mentions humaines et non-humaines. L'étude des annotations de référence du corpus d'entraînement nous a montré que la quasi totalité des mentions humaines sont des patients ou des médecins. Nous avons donc conçu des tests pour identifer les mentions qui réfèrent au patient ou au médecin. Une instance spécifique du médecin est l'auteur du rapport, que nous avons également traitée séparément. Cette méthode s'apparente au modèle dit « entité-mention » [6], dans lequel on cherche à relier chaque mention à une entité en cours de constitution et représentée par un groupe de mentions antérieures. Dans notre cas, les entités sont déterminées a priori par étude du corpus.

Pour détecter ces « entités connues », nous avons conçu des patrons reposant sur des indices internes (qui s'appliquent à la mention) et externes (qui s'appliquent aux contextes gauche et droit de la mention). Il s'agit d'expressions régulières (celles du langage Perl) fondées sur les caractères. Ces patrons ont eux-mêmes été préalablement identifiés dans la partie entraînement du corpus ODIE en combinant l'étude des chaînes de coréférence fournies avec le corpus et l'étude des occurrences des mentions dans les textes.

Ainsi, une partie importante des chaînes de coréférence de type humain ont pu être identifiées comme référant au patient (présence dans une des mentions du mot *patient* ou d'une de ses abréviations) ou à un médecin (présence dans une des mentions du titre *dr.* ou *m.d.*). Cela nous a donné des exemples d'ensembles de mentions de type patient ou médecin.

Nous avons également calculé la distribution de fréquence des mentions de type humain dans le corpus d'entraînement. Pour les plus fréquentes d'entre elles, nous avons pu déterminer, en nous référant au résultat précédent, si elles référaient au patient ou à un médecin.

Nous avons alors créé des patrons pour couvrir ces occurrences les plus fréquentes, en généralisant à la main les expressions observées pour que les patrons s'appliquent à des familles de cas. Les expressions retenues pour identifier qu'une mention de type humain désigne le patient incluent : la présence du mot *patient* ou de son abréviation *pt* ; débuter par le titre *mr*, *mrs*, *ms* ; une désignation par le sexe ou l'ethnie (*male, man, gentleman, female, woman, caucasian*), l'âge (*year-old*) ou une classe d'âge (*infant, baby, newborn, toddler, boy, girl*), le métier (*student, secretary*). Pour le médecin, le titre (*m.d., dr*), la fonction (*doctor, attending, cardiologist, internist, nephrologist, neurologist, pcp, primary care physician, physician, primary cardiologist, primary care doctor, primary care provider, primary doctor, psychiatrist, surgeon*) ou la spécialité du service (*cardiology, hematology*) ont été inclus. Enfin, pour l'auteur du compte rendu, la mention devait être l'un des pronoms *i, my, myself, we, our, us*. Pour illustrer le format em-

5. Dans le système de base, nous avons simplement défini la tête comme étant le dernier mot de l'entité.

ployé pour fournir ces expressions régulières, nous montrons ci-dessous la règle pour détecter une référence à l'auteur. Une mention de l'un des types séparés par des ' :' désigne l'auteur si le patron s'applique à cette mention :

```
people:pronoun:none:other → author
if ^(i|my|myself|we|our|us)$
```

Considérer les pronoms *he* et *she* comme systématiquement du type patient réduisait significativement les performances (environ 3 à 4 points sur la F-mesure moyenne) sur plusieurs de nos corpus d'entraînement, nous ne les avons donc pas intégrés dans les patrons de l'entité patient.
Les indices externes sont nécessaires pour catégoriser les pronoms (*he*, *she*, *his*, *her*). Certains de nos patrons externes sont fondés sur des verbes qui occurrent plus particulièrement avec des patients ou des médecins comme sujet. Nous avons en effet recherché dans le corpus d'entraînement les occurrences des pronoms *he*/*she* suivies d'un verbe, et avons observé la prévalence de certaines formes (passif, etc.) en association avec le patient, l'examen des chaînes de coréférence associées au corpus de développement fournissant le test « officiel » de la référence du pronom. Nous avons alors raffiné progressivement ces expressions régulières pour détecter ces formes. Nous avons également trouvé que les constructions passives (*have*|*was* + past perfect) sont fortement associées au patient, de même qu'une série de verbes d'état ou de changement d'état (*began*, *underwent*) et (moins fortement) les verbes des temps du passé. La présence d'un adverbe (*first*, *then*, *never*, *still*, *now*) entre le sujet *he*/*she* et le verbe renvoie également au patient. Inversement, les médecins effectuent, prescrivent, recommandent, suggèrent, etc., souvent au temps présent. D'autres patrons mettant en jeu les expressions *signed by* author et *his/her office*, *secretary*, *fax*, *opinion*, *impression*, *note*, *schedule* (doctor) renvoient au médecin. Un type supplémentaire de patron tire profit de l'inclusion hiérarchique d'une mention typée dans une autre : dans la mention *his/her X*, *his/her* renvoie à l'entité patient (*his chest*) sauf quand *(his) X* renvoie au patient (*his patient*) (entité doctor).
Les mentions non-humaines désignent des concepts médicaux. Le Metathesaurus de l'UMLS [13] recense plusieurs millions de concepts médicaux différents, pour lesquels il fournit plus de cinq millions de libellés anglais différents. Nous avons donc cherché à représenter ces mentions par des concepts de l'UMLS (les pronoms et le type none du corpus ODIE n'ont pas fait l'objet de ce type de traitement). De plus, le système MetaMap (`http://metamap.nlm.nih.gov/`) permet de détecter dans un texte les libellés anglais de l'UMLS ou certaines de leurs variantes. Nous avons donc appliqué MetaMap sur chaque mention. Nous avons restreint MetaMap aux types sémantiques de l'UMLS pertinents pour le type i2b2/ODIE de chaque mention. Nous avons également utilisé la fonctionnalité de désambiguïsation sémantique de MetaMap, ce qui augmente la proportion de mentions pour lesquelles un seul concept de l'UMLS est renvoyé. Parmi les critères de MetaMap pour sélectionner les concepts figure la « centralité », autrement dit, intégrer la tête du syntagme d'entrée. Il s'agit d'un point important puisque la plupart du temps, en examinant les mentions candidates à la coréférence, une représentation identique de leurs têtes implique qu'elles sont dans l'une des relations sémantiques préférées pour produire une coréférence : identité ou relation spécifique/générique.

Plus de règles. Sur la base de ces représentations, nous avons affiné les trois règles précédemment décrites et en avons défini de nouvelles. Des exceptions à *String_Equality* ont été introduites pour les pronoms que nous ne pouvions inclure dans de futures règles : *this, that, these, they, there, it, its*. Nous n'avons pas bloqué les coréférences sur *he/he* et *she/she* car cela réduisait significativement la performance (environ 5 points sur la F-mesure moyenne) sur plusieurs de nos corpus d'entraînement.
La représentation conceptuelle des mentions a été utilisée dans les règles qui suivent. *Common_Concept* : chaque mention a une représentation conceptuelle et elles partagent un concept de l'UMLS. Au-delà du bon cas où les concepts UMLS couvrent l'ensemble des mots des mentions, cette règle détecte le plus souvent des mentions dans une relation spécifique/générique. *Same_Entity* : les deux mentions réfèrent à la même entité connue. Des contraintes supplémentaires ont été définies pour l'entité Doctor : pour tester qu'elles ont le même nom, deux entités « médecin » doivent également avoir une similarité suffisamment élevée (empiriquement établie à 0,5 sur le corpus d'entraînement). Cependant, ceci écarte les coréférences entre un nom de médecin (*Dr Jones*) et sa fonction (*cardiologist*).

6 Évaluation et discussion

Des expérimentations pour évaluer nos deux systèmes ont été réalisées sur les données de développement de chaque corpus sus-mentionné. Le modèle pour la détection des singletons a utilisé l'implémentation SVM-light [19] de SVM. Le paramètre de marge souple pour SVM a été optimisé sur les données de développement, alors que le paramètre affectant le poids des erreurs sur des exemples positifs a été fixé d'après le ratio d'exemples positifs et négatifs, comme suggéré par les auteurs de SVM-light.
Les mesures utilisées pour évaluer les performances des systèmes dans le défi i2b2/VA (rappel, précision et F-mesure selon B^3, MUC, BLANC et CEAF, voir [20]) ont des comportements différents et n'étaient pas toujours intuitives. Nous présentons le résultat en utilisant ces métriques officielles.
Le tableau 3 indique la F-mesure obtenue sur chaque sous-corpus du corpus d'entraînement global ODIE par nos différentes méthodes (pour plus de clarté, nous ne mentionnons pas les valeurs de rappel et précision). Les résultats varient de manière significative selon la métrique. Sur chaque sous-corpus, nous avons obtenu nos meilleurs résultats en utilisant le système déterministe UM. Ceci s'inscrit dans la continuité des travaux du domaine [4, 5]. Pré-

TAB. 3 – F-mesure pour nos trois systèmes de coréférence sur les six sous-corpus d'entraînement ODIE. SD = détection des singletons (étape 1+étape 2a), 3R = 3 règles (étape 2b), UM = règles et UMLS (étape 2c)

Métriques	Mayo Clinic						Pittsburgh											
	Clinical			Pathology			Discharge			Other			Radiology			Surgery		
	SD	3R	UM	SD	3R	UM	SD	3R	UM	SD	3R	UM	SD	3R	UM	SD	3R	UM
B^3	0,83	0,88	0,91	0,83	0,86	0,88	0,86	0,88	0,92	0,89	0,95	0,93	—	0,83	0,88	—	0,74	0,84
MUC	0,16	0,75	0,84	0,14	0,30	0,55	0,14	0,67	0,82	0,18	0,81	0,88	—	0,37	0,64	—	0,53	0,81
Blanc	0,52	0,75	0,71	0,56	0,61	0,71	0,52	0,66	0,76	0,51	0,71	0,76	—	0,59	0,69	—	0,65	0,76
CEAF	0,37	0,61	0,82	0,53	0,60	0,67	0,42	0,61	0,72	0,40	0,63	0,68	—	0,46	0,54	—	0,47	0,65
Moyenne	0,45	0,74	0,81	0,50	0,59	0,70	0,47	0,72	0,92	0,49	0,78	0,83	—	0,55	0,69	—	0,58	0,77

cisons toutefois que pour le système SD, nous n'avons pas pu profiter d'un modèle d'apprentissage pour les sous-corpus Pittsburgh ; à la place, nous avons appliqué un modèle construit sur le sous-corpus Clinical sur tous les sous-corpus de Pittsburgh.

Lors de l'évaluation officielle de la campagne i2b2/VA, deux nouvelles parties des corpus ODIE et i2b2 nous ont été fournies, avec leurs mentions annotées. Nous avons fait tourner diverses versions de nos systèmes sur ces corpus et avons renvoyé les chaînes de coréférence calculées pour ces mentions. La F-mesure moyenne sur toutes les métriques sur tous les sous-corpus a été utilisée pour le classement. Notre meilleur système (UM) s'est classé en position médiane : 4e sur 8 sur le corpus ODIE (F = 0,802 ; moyenne des participants : 0,734, médiane 0,800, max 0,827, min 0,417), et 9e sur 16 sur le corpus i2b2 (F = 0,856, moyenne 0,844, médiane 0,859, max 0,915, min 0,559). Une piste supplémentaire concernait un système complet, qui partait de textes bruts dont les mentions n'étaient pas préannotées. Nous faisons partie des rares équipes qui ont pu mettre en place un tel système, qui s'est classé premier sur les trois participants à cette piste (F = 0,729, médiane 0,699, min 0,417). Notre système UM était précédé dans ce cas de notre système de reconnaissance de mentions, non détaillé ici.

Après soumission de nos résultats, nous avons reçu les annotations de référence pour les chaînes de coréférence de ces corpus. Cela nous a permis d'observer l'impact sur les performances des divers éléments décrits plus haut dans l'étape 2c. Les résultats de plusieurs versions du système à base de règles sont indiqués dans le tableau 4 : 3R est le système à trois règles (Étape 2b) ; UM-0 (Étape 2c) est la version à laquelle est ajoutée la représentation par des concepts de l'UMLS et la détermination d'« entités connues » par des indices internes, ainsi que le blocage des coréférences par égalité de forme sur les pronoms *this, that, it* ; UM-I ajoute les patrons par Inclusion hiérarchique et le blocage de *these, they, there, its* ; et UM-T ajoute les patrons par indices externes, qui reviennent au Texte. Comme souvent, le gain se réduit au fil des ajouts, mais il reste encore important (1,3 puis 0,5 points).

TAB. 4 – F-mesure moyenne sur les quatre métriques et les six sous-corpus d'évaluation ODIE

Variante	3R	UM-0	UM-I	UM-T
F-mesure moyenne (%)	58,9	76,2	77,5	78,0

Une analyse plus fine des résultats nous a permis de voir que la marge de progression restante est encore importante. Elle nous a aussi montré que le gain obtenu est très dépendant des sous-corpus : certains comprennent beaucoup de références à des personnes (patient, médecins, etc.), et bénéficient fortement des améliorations apportées au traitement des pronoms et des entités connues, alors que d'autres en comportent peu et voient leurs résultats inchangés.

On peut se poser la question de la généralité de l'approche et des règles mises en place. De fait, les règles *Who_Ref* et *Which_Ref* devraient s'appliquer correctement à d'autres domaines, la règle des chaînes identiques aussi. En revanche, les règles de détection du patient et des médecins sont clairement spécifiques au domaine et au genre de textes abordé, puisqu'elles reposent sur la détection d'« entités connues » dans ces textes. Par ailleurs, d'autres participants au défi i2b2/VA 2011 ont constaté expérimentalement qu'en déconnectant les connaissances spécifiques au domaine présentes dans leurs systèmes, les performances baissaient seulement de quelques points de F-mesure.

Enfin, nous avons constaté que d'autres participants au défi ont mis en place des méthodes similaires à celle qui vise à catégoriser les personnes (patient, médecins, famille) ; certains ont conçu des patrons utilisant comme nous les verbes ; d'autres ont eu recours à un apprentissage supervisé s'appuyant sur les mots proches des pronoms.

7 Conclusion

Nous avons présenté les expérimentations que nous avons réalisées pour détecter les coréférences de concepts dans des documents médicaux. Nous avons testé à la fois une approche par apprentissage reposant sur les SVM et une approche à base de règles qui effectue une identification, à base de connaissances, d'entités connues et de concepts, et traite des cas limités d'anaphores pronominales. Notre meilleure méthode est celle à base de règles, qui obtient une F-mesure moyenne qui varie de 0,69 à 0,92 sur le sous-corpus ODIE.

Nous n'avons testé qu'un petit sous-ensemble des combinaisons possibles de nos modules individuels, et plusieurs directions méritent d'être explorées. En premier lieu, le module de détection des singletons a uniquement été testé avec un classifieur simple de paires de mentions (SD) : nous devons également le tester avec notre meilleur détecteur de paires de mentions (UM). En second lieu, une série de règles simples à ajouter mérite d'être testée pour couvrir plus de cas d'anaphores pronominales. De plus, d'autres cas de relations générique/spécifique entre mentions doivent être testées au travers des relations du Metathesaurus de l'UMLS. Enfin, nous n'avons que faiblement utilisé les structures de discours et nous sommes principalement reposés sur la compatibilité sémantique des mentions de concepts. L'état actuel de notre meilleur système devrait fournir un bon point de départ pour tester quels filtres fondés sur le discours peuvent apporter des améliorations notables.

Remerciements Ce travail a été partiellement réalisé dans le cadre des projets DoXa (financement CapDigital) et Quaero (financement Oséo, agence française pour l'innovation et la recherche).

Références

[1] J. R. Hobbs, "Resolving pronoun references," *Lingua*, vol. 44, pp. 311–338, 1978.

[2] S. Lappin and H. J. Leass, "An algorithm for pronominal resolution," *Computational Linguistics*, vol. 20, no. 4, pp. 535–561, 1994.

[3] A. Rahman and V. Ng, "Supervised models for coreference resolution," in *Proc. EMNLP 2009*, (Singapore), pp. 968–977, Association for Computational Linguistics, 2009.

[4] A. Haghighi and D. Klein, "Simple coreference resolution with rich syntactic and semantic features," in *Proc. EMNLP 2009*, (Singapore), pp. 1152–1161, Association for Computational Linguistics, August 2009.

[5] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and C. Manning, "A multi-pass sieve for coreference resolution," in *Proc. EMNLP 2010*, (Cambridge, MA), pp. 492–501, Association for Computational Linguistics, 2010.

[6] J. Zheng, W. W. Chapman, R. S. Crowley, and G. K. Savova, "Coreference resolution : A review of general methodologies and applications in the clinical domain," *J Biomed Inform*, vol. 44, pp. 1113–1122, Dec. 2011.

[7] O. Uzuner, I. Solti, and E. Cadag, "Extracting medication information from clinical text," *J Am Med Inform Assoc*, vol. 17, no. 5, pp. 514–518, 2010.

[8] O. Uzuner, B. R. South, S. Shen, and S. L. Duvall, "2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text," *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 552–556, 2011.

[9] C. Grouin, M. Dinarelli, S. Rosset, G. Wisniewski, and P. Zweigenbaum, "Coreference resolution in clinical reports — the LIMSI participation in the i2b2/VA 2011 challenge," in *i2b2/VA 2011 Challenge Workshop* (Özlem Uzuner et al., ed.), (Washington, DC), i2b2, 2011. 10 pages.

[10] J.-F. McCarthy and W.-G. Lehnert, "Using decision trees for coreference resolution," in *Proceedings of IJCAI'95*, (Montréal, Canada), 1995.

[11] W. M. Soon, H. T. Ng, and D. C. T. Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational Linguistics*, vol. 27, no. 4, pp. 521–544, 2001.

[12] T. Y. He, "Coreference resolution on entities and events for hospital discharge summaries," Master's thesis, Massachusetts Institute of Technology, 2007.

[13] D. A. Lindberg, B. L. Humphreys, and A. T. McRay, "The Unified Medical Language System," *Meth Inform Med*, vol. 32, no. 4, pp. 281–291, 1993.

[14] O. Uzuner, J. Pestian, and B. South, "The i2b2/va 2011 challenge," in *i2b2 Workshop Proceedings*, (Washington, DC), 2011.

[15] A.-L. Minard, A.-L. Ligozat, A. Ben Abacha, D. Bernhard, B. Cartoni, L. Deléger, B. Grau, S. Rosset, P. Zweigenbaum, and C. Grouin, "Hybrid methods for improving information access in clinical documents : Concept, assertion, and relation identification," *J Am Med Inform Assoc*, vol. 18, no. 5, pp. 588–593, 2011. Published Online First : 19 May 2011.

[16] O. Galibert, *Approches et méthodologies pour la réponse automatique à des questions adaptées à un cadre interactif en domaine ouvert*. PhD thesis, Université Paris Sud, Orsay, 2009.

[17] S. Rosset, O. Galibert, G. Bernard, E. Bilinski, and G. Adda, "The LIMSI multilingual, multitask QAst system," in *Proc. CLEF 2008*, (Berlin, Heidelberg), pp. 480–487, Springer-Verlag, 2009.

[18] Z. Guodong, S. Jian, Z. Jie, and Z. Min, "Exploring various knowledge in relation extraction," in *ACL'05 : Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, (Morristown, NJ, USA), pp. 427–434, Association for Computational Linguistics, 2005.

[19] T. Joachims, "Training linear SVMs in linear time," in *KDD'06 : Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (New York, NY, USA), pp. 217–226, ACM, 2006.

[20] M. Recasens, L. Màrquez, E. Sapena, M. A. Martí, M. Taulé, V. Hoste, M. Poesio, and Y. Versley, "SemEval-2010 Task 1 : Coreference resolution in multiple languages," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, (Uppsala, Sweden), pp. 1–8, Association for Computational Linguistics, July 2010.