Détection de concepts et granularité de l'annotation

Pierre Zweigenbaum[1] Thomas Lavergne[2]
(1) LIMSI, CNRS, Université Paris-Saclay, 91405 Orsay, France
(2) LIMSI, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91405 Orsay, France
pz@limsi.fr, lavergne@limsi.fr

RÉSUMÉ

Nous nous intéressons ici à une tâche de détection de concepts dans des textes sans exigence particulière de passage par une phase de détection d'entités avec leurs frontières. Il s'agit donc d'une tâche de catégorisation de textes multiétiquette, avec des jeux de données annotés au niveau des textes entiers. Nous faisons l'hypothèse qu'une annotation à un niveau de granularité plus fin, typiquement au niveau de l'énoncé, devrait améliorer la performance d'un détecteur automatique entraîné sur ces données. Nous examinons cette hypothèse dans le cas de textes courts particuliers : des certificats de décès où l'on cherche à reconnaître des diagnostics, avec des jeux de données initialement annotés au niveau du certificat entier. Nous constatons qu'une annotation au niveau de la « ligne » améliore effectivement les résultats, mais aussi que le simple fait d'appliquer au niveau de la ligne un classifieur entraîné au niveau du texte est déjà une source d'amélioration.

ABSTRACT

Concept detection and annotation granularity

We address here the detection of concepts in texts with no requirement for first spotting entities and their boundaries. This is thus a multilabel text categorization task, with a dataset with text-level annotations. We hypothesize that a finer-grained annotation, typically at the sentence level, should improve the performance of an automatic concept detector trained on these data. We examine this hypothesis in the case of specific short texts : death certificates in which we aim to detect diagnoses, with datasets initially annotated at the level of full certificates. We observe that a "line"-level annotation does improve concept detection ; but also that simply applying at the sentence level a classifier trained at the text level is already a source of improvement.

MOTS-CLÉS : Extraction d'information ; catégorisation de textes ; détection de concepts ; granularité de l'annotation ; diagnostics.

KEYWORDS: Information extraction; text categorization; concept detection; annotation granularity.

1 Introduction

La détection de concepts dans des textes consiste à y relever la présence de concepts d'un référentiel termino-ontologique. Citons par exemple *Absence acquise de poumon* (Z90.2 dans la CIM-10, Classification internationale des maladies, version 10, de l'OMS (OMS, 1993)) ou encore *Pneumonies, sans précision* (J18.9 dans la CIM-10). Ces concepts fournissent une représentation des informations utiles dans divers systèmes d'information (Zweigenbaum, 2017). Par exemple, dans le domaine médical, les « descripteurs » du thésaurus MeSH sont employés pour l'indexation et la recherche de

notices bibliographies d'articles scientifiques dans la base MEDLINE, et les classes de la Classification internationale des maladies sont employées pour caractériser les séjours hospitaliers à des fins médico-économiques ou pour compiler à l'échelle nationale et internationale les statistiques de mortalité à partir des textes des certificats de décès.

La détection de concepts est généralement abordée comme une étape suivant la détection d'entités nommées (Nouvel *et al.*, 2015). Il s'agit alors d'une tâche de normalisation d'entités, aussi appelée plus récemment liage d'entités (Zheng *et al.*, 2015). Cette tâche demande de déterminer, pour chaque expression décrivant une entité, déjà définie par ses frontières et généralement un type, le concept le plus approprié parmi ceux d'un référentiel donné (typiquement une terminologie, ontologie ou base de données). Les corpus annotés indiquent donc pour chaque mention d'une entité le concept attendu pour celle-ci.

Nous nous plaçons dans un contexte différent où la détection de concepts est définie comme une tâche de catégorisation de textes. Étant donné un texte, sans indication préalable d'entités et de leurs frontières, il s'agit de déterminer les concepts mentionnés dans ce texte. Les corpus annotés fournissent alors simplement la liste des concepts associés à chaque texte, sans indication de localisation. D'un côté, supprimer l'évaluation des frontières simplifie la tâche. D'un autre côté, l'absence d'indication sur la position des concepts dans les annotations fournies rend la tâche plus difficile : on ne sait pas a priori quels mots sont révélateurs des concepts à trouver, ce qui donne moins d'indices pour l'entraînement. De plus, par rapport aux tâches courantes de catégorisation de textes, qui considèrent quelques classes (Sebastiani, 2002), la détection de concepts s'attaque souvent à des milliers, voire des dizaines de milliers de classes (Tsatsaronis *et al.*, 2015) : de l'ordre de 13 000 dans la CIM-10 et de 27 000 dans MeSH. Parmi les types de méthodes employées pour aborder cette tâche, les deux plus fréquentes sont celles fondées sur des dictionnaires ou terminologies existants (Aronson & Lang, 2010), et celles fonctionnant par apprentissage supervisé (Lin & Wilbur, 2007).

Nous nous intéressons dans cet article à la granularité des annotations que l'on peut employer pour détecter des concepts dans cette situation. Par rapport à un corpus annoté au niveau des textes, un corpus annoté à un niveau plus fin, par exemple des énoncés, aide-t-il à mieux détecter les concepts de chaque texte ? Nous partons d'un corpus avec deux niveaux d'annotation et comparons les performances obtenues dessus en détection de concepts pour tester les deux hypothèses suivantes :

— Lorsque l'on utilise des méthodes par apprentissage supervisé, leur entraînement sur une annotation au niveau de granularité plus fin devrait améliorer la détection de concepts par rapport à leur entraînement sur une annotation au niveau des textes.
— Les méthodes à base de dictionnaire ne devraient pas être affectées par ces différences d'annotation.

2 Données et méthodes

2.1 Cas d'usage et données : codage CIM-10 dans des certificats de décès

Nous utilisons dans ces expériences les données d'entraînement de la tâche CLEF eHealth 2017 (`https://sites.google.com/site/clefehealth2017/task-1`), qui fournissent des annotations aux deux niveaux recherchés. Ces données concernent des certificats de décès français, textes courts auxquels sont associés des concepts désignant les maladies ou événements ayant mené au décès (tableau 1). Ces concepts sont choisis parmi les quelque 13 000 classes de la Classification internatio-

Grain	Segment annoté	Concepts
ligne	pneumopathie nosocomiale	J18.9
ligne	insuffisance respiratoire chronique	J96.1
ligne	insuffisance cardiaque chronique, pneumonectomisé pour cancer pulmonaire, lymphome non hodgkinien en abstention thérapeutique	C85.9, I50.9, C34.9, Z90.2
certif.	pneumopathie nosocomiale XYZT insuffisance respiratoire chronique XYZT insuffisance cardiaque chronique, pneumonectomisé pour cancer pulmonaire, lymphome non hodgkinien en abstention thérapeutique	J18.9, J96.1, I50.9, Z90.2, C34.9, C85.9

TABLE 1 – Exemple de certificat (*certif.*) et de ses trois énoncés (« lignes »)

nale des maladies, version 10 (codes CIM-10, (OMS, 1993)). Un certificat se compose de un à cinq énoncés, un par ligne dans le formulaire officiel, chaque énoncé rapportant un ou plusieurs maladies ou autres événements. On constate que ces énoncés sont essentiellement des listes de maladies, qui n'emploient naturellement pas nécessairement les mêmes termes que les libellés des classes de la CIM-10, comme le montre l'exemple de J18.9 et Z90.2, présents dans les première et troisième ligne du tableau 1, dont les libellés sont donnés dans l'introduction.

	Total		Entraînement		Test	
	certificat	ligne	certificat	ligne	certificat	ligne
Certificats	64 980	64 980	61 521	61 521	3 459	3 459
Segments annotés	64 980	192 641	61 521	182 643	3 459	9 998
Concepts	262 809	257 772	249 249	244 474	13 560	13 298

TABLE 2 – Jeux de données comparables avec des annotations au niveau du certificat ou de la ligne

La version d'origine du corpus contient des annotations au niveau des certificats entiers (ligne *certif.* du tableau 1 et colonnes *certificat* du tableau 2), que les organisateurs (Lavergne *et al.*, 2016) ont reportées au niveau des lignes (lignes ou colonnes *ligne* des tableaux) [1]. Le tableau 2 donne les statistiques des corpus résultants, contenant près de 65 000 certificats composés de près de 193 000 lignes, que nous avons divisés en deux parties pour nos expériences : entraînement ($\sim 95\ \%$) et test ($\sim 5\ \%$). Chaque partie est donc annotée au niveau du certificat ou de la ligne (rangée *Segments annotés*). On constate que le nombre de concepts (rangée *Concepts*) est légèrement inférieur ($\sim -2\ \%$) dans la version *ligne* : certains concepts se sont perdus lors du report (semi-automatique) des certificats aux lignes, ce qui pourra légèrement défavoriser les traitements effectués à partir de ces lignes. Le corpus total contient 3 290 concepts différents en version *certificat* et 3 220 en version *ligne*, ce qui va dans le même sens.

Les organisateurs fournissent aussi des dictionnaires employés par le CépiDc [2] dans le processus réel de codage des certificats de décès. Nous utilisons ici celui de 2011 (les résultats sont similaires avec celui de 2012, également fourni ; les certificats traités vont de 2006 à 2012). Il contient 156 937 entrées pour 6 093 concepts différents. Ce dictionnaire contient tous les concepts mentionnés dans le corpus sauf 249 dans la version *certificat* et 245 dans la version *ligne* (7,6 % des concepts différents).

En préalable à tout traitement, nous supprimons les mots outils et les ponctuations, racinisons, passons

1. Constatant que certains certificats s'étaient perdus dans la version annotée au niveau des certificats, nous les avons aussi supprimés de la version *ligne*.

2. Centre d'épidémiologie des causes médicales de décès, Unité Inserm US10, `http://www.cepidc.inserm.fr/`.

en minuscules et supprimons les diacritiques des mots restants dans chaque ligne. Dans les données par certificat, toutes les lignes d'un même certificat sont d'abord regroupées en un texte et séparées par un symbole arbitraire (séquence *XYZT* absente du corpus) servant de séparateur d'énoncé.

Méthode à base de dictionnaire La projection de dictionnaire consiste à rechercher dans les textes des occurrences de termes connus associés à des concepts. Étant donné une liste de concepts avec pour chacun un ou plusieurs termes, nous en effectuons une détection efficace (gauche-droite, plus longue séquence) en encodant le dictionnaire dans un transducteur à états finis. Le même traitement initial qu'aux textes est appliqué aux entrées de dictionnaire. De ce fait, la recherche exacte d'une séquence de mots ainsi normalisés fournit une recherche souple de termes, robuste à des variations de casse, de désinence, de diacritiques et de mots outils. Le dictionnaire peut proposer plusieurs concepts pour la même expression, qui est alors ambiguë. Dans ce cas nous testons deux méthodes : *tous* renvoie tous les concepts associés, *nonamb* ne renvoie pas de concept si l'expression est ambiguë.

Une méthode hybride : calibrage du dictionnaire Certains des concepts détectés par le dictionnaire peuvent être erronés. Pour réduire ce type d'erreurs, nous entraînons un classifieur binaire (OUI/NON) (un SVM) qui décide si un concept est pertinent ou pas. L'entraînement est effectué sur notre jeu d'entraînement, en prenant comme attributs le concept proposé par le dictionnaire et le sac de mots de la ligne ou du texte d'entrée. Lors de l'application au jeu de test, le classifieur entraîné est appliqué à chaque concept proposé par le dictionnaire. S'il le classe comme un NON, nous supprimons ce concept de la sortie produite.

Détection de concepts par apprentissage supervisé Nous entraînons un classifieur multiclasse pour décider des concepts associés à un texte ou une ligne d'entrée. Après des expériences initiales sur notre jeu d'entraînement, nous avons opté pour les attributs suivants : (i) le sac de mots de l'entrée, normalisé comme expliqué plus haut ; (ii) le sac de trigrammes de caractères de l'entrée, également calculé après normalisation des mots ; cela fournit de la robustesse aux fautes d'orthographe et autres variantes terminologiques ; (iii) l'année de codage du certificat : cela permet de prendre en compte des variations de règles de codage au fil des années. Après divers essais réalisés en validation croisée sur le jeu d'entraînement, nous avons constaté que ces attributs fournissaient les meilleurs résultats, et que la plupart du temps un SVM linéaire était meilleur que plusieurs autres classifieurs (régression logistique, forêt aléatoire, k plus proches voisins, etc.).

Nos expériences initiales se sont réalisées sur des classifieurs monoétiquette, qui ne produisent qu'un concept pour chaque entrée. Cependant, comme dans la situation réelle chaque entrée peut mener à plusieurs concepts, cette solution n'est pas optimale. Nous avons donc également entraîné un classifieur multiétiquette, beaucoup plus gourmand computationnellement, car il entraîne un classifieur par concept. Par exemple, l'implémentation que nous avons utilisée (scikit-learn, LinearSVM, métaclassifieur OneVsRestClassifier) a pris trois heures pour l'entraînement (182 643 exemples, de l'ordre de 3 000 classes) et 25 minutes pour le test dans la version *ligne* du corpus (9 998 exemples).

Les méthodes présentées peuvent sembler simples ; en pratique, l'union du dictionnaire calibré et du SVM mono-étiquette détient les meilleurs résultats publiés (Zweigenbaum & Lavergne, 2016) hors soumissions officielles sur le corpus de test de CLEF eHealth 2016 (Névéol *et al.*, 2016), et le SVM multi-étiquette seul ou combiné au dictionnaire a obtenu les meilleurs résultats hors soumissions officielles sur les corpus de test français (ligne et certificat) de CLEF eHealth 2017 (Zweigenbaum & Lavergne, 2017).

Granularité de l'entraînement, de l'application et de l'évaluation Pour l'entraînement d'un classifieur ou le calibrage du dictionnaire, nous disposons d'annotations au niveau des certificats

entiers ou de leurs lignes (voir les tableaux 1 et 2). C'est le premier paramètre (*entraînement*) que nous faisons varier dans nos expériences. Nous pouvons ensuite appliquer la détection de concepts au niveau d'un certificat entier ou de chaque ligne d'un certificat. C'est notre deuxième paramètre (*test*). Enfin, dans le cas où l'application se fait sur chaque ligne, l'évaluation peut se faire par comparaison à la référence de chaque ligne, ou en regroupant d'abord les concepts proposés pour toutes les lignes d'un certificat et en les comparant à la référence du certificat. C'est notre troisième paramètre (*éval*).

3 Résultats

La simple projection de dictionnaire (tableau 3, rangées avec – dans la colonne *Cal*) ne change pas avec le jeu d'entraînement. En revanche, l'évaluation au niveau de chaque ligne de certificat est plus exigeante que celle effectuée au niveau des certificats entiers : un concept peut être incorrect pour une ligne mais s'avérer correct pour le certificat du fait qu'il apparaît dans une autre ligne du même certificat. On le constate pour le dictionnaire brut, mais aussi pour ses variantes avec calibrage (rangées avec *c-t* dans la colonne *Cal*). Le calibrage sur les certificats a des résultats similaires à celui sur les lignes. Il apporte dans les deux cas une grande amélioration de la précision et de la F-mesure. Sa combinaison à la méthode *nonamb* est encore plus précise, mais un bien meilleur rappel est obtenu par la méthode *tous*, dont le calibrage apporte la meilleure F-mesure.

			test=certificat					test=ligne				
Entr.	Méth.	Cal.	Sys.	TP	P	R	F1	Sys.	TP	P	R	F1
certif	nonamb	–	10 812	8 863	82,0	65,4	72,7	10 866	8 677	79,9	65,3	71,8
certif	nonamb	c-t	8 927	8 465	**94,8**	62,4	75,3	8 983	8 347	**92,9**	62,8	74,9
certif	tous	–	18 466	10 690	57,9	**78,8**	66,8	18 591	10 450	56,2	**78,6**	65,5
certif	tous	c-t	10 711	9 787	91,4	72,2	**80,7**	10 774	9 640	89,5	72,5	**80,1**
ligne	nonamb	–	10 812	8 863	82,0	65,4	72,7	10 867	8 677	79,9	65,3	71,8
ligne	nonamb	c-t	8 368	7 999	**95,6**	59,0	73,0	8 843	8 293	**93,8**	62,4	74,9
ligne	tous	–	18 466	10 690	57,9	**78,8**	66,8	18 590	10 450	56,2	**78,6**	65,5
ligne	tous	c-t	9 949	9 196	92,4	67,8	78,2	10 548	9 575	90,8	72,0	**80,3**

TABLE 3 – Calibrage du dictionnaire sur des certificats (13 560 concepts) et des lignes (13 298 concepts). Le gras indique les meilleurs résultats pour la précision (P), le rappel (R) et la F-mesure (F1) (en pourcentages). Entr. = entraînement. Méth. = méthode du dictionnaire (voir le texte). Cal. = Type de calibrage : – = aucun, c-t = concept+mots. Sys. = nb de concepts prédits. TP = vrais positifs.

Les expériences réalisées en classification supervisée confirment que si l'objectif est de détecter les concepts de chaque énoncé-ligne (colonnes *éval=ligne* en partie droite du tableau 4) plutôt qu'au niveau du texte-certificat entier, il vaut mieux disposer d'une annotation au niveau de chaque énoncé. Ce point est très clair en classification monoétiquette (colonne éval = ligne, +18pt de F-mesure de 59 à 77 %) ; dans notre jeu de données, la classification multiétiquette est nettement plus robuste et ne perd que 2 points de F-mesure (de 87 à 85 %) du fait d'une diminution de précision (-5 pt de 88 à 83 %) à rappel relativement constant.

Si l'objectif est de détecter globalement les concepts présents dans le texte (colonnes éval = certificat du tableau 4), ces expériences montrent qu'il reste intéressant de passer par l'énoncé de deux façons :

— Un effort d'annotation des énoncés est le plus payant. En classification monoétiquette, entraîner un classifieur sur les énoncés ainsi annotés apporte un très large gain par rapport à un

		test=certificat éval=certificat			test=ligne éval=certificat ←			éval=ligne		
Entr.	Méth.	P	R	F1	P	R	F1 ←	P	R	F1
certif	mono	**85,5**	21,8	34,7	70,7	**51,6**	**59,6** ←	68,8	51,7	59,0
ligne	mono	80,7	20,6	32,8	**91,4**	67,2	**77,5** ←	89,9	**67,6**	77,1
certif	multi	85,4	79,9	82,5	**85,7**	**86,7**	**86,2** ←	83,4	86,5	84,9
ligne	multi	85,4	61,3	71,4	**90,3**	**86,3**	**88,2** ←	88,0	86,2	87,1

TABLE 4 – Classification supervisée (SVM linéaire : LinearSVC de scikit learn) sur les certificats et sur les lignes. Le gras note les meilleurs résultats par rangée en précision (P), rappel (R) et F-mesure (F1) (en pourcentages).

entraînement et test sur les textes (rangée ligne / mono, colonne test=ligne, éval=certificat : +42 points de F-mesure de 35 à 77 %).
De même, en classification multiétiquette, l'entraînement sur des lignes annotées a des résultats supérieurs à l'entraînement et test sur les textes (+5 points de F-mesure de 83 à 88 %).

— Même sans disposer d'annotations au niveau des énoncés, entraîner un classifieur monoétiquette sur les textes annotés puis l'appliquer individuellement sur les énoncés augmente mécaniquement le nombre de concepts proposés et ainsi le rappel, au prix d'une baisse de précision (rangée certif / mono, colonne test=ligne, éval=certificat : +30pt de rappel de 22 à 52 %, +25 points de F-mesure de 35 à 60 %). Le rappel est en effet limité par la classification monoétiquette, qui ne propose qu'un concept par unité textuelle.
Bien que n'ayant pas cette limitation en rappel, le classifieur multiétiquette bénéficie lui aussi d'une application sur chaque énoncé plutôt que sur un texte entier : il augmente un peu sa précision et fortement son rappel (+7 pt de rappel de 80 à 87 %, +3 pt de F-mesure de 83 à 86 %). Nous supposons que c'est lié à deux phénomènes. D'une part, la représentation choisie étant en sac de mots, l'application à des énoncés individuels réduit le risque de confusion dû à des mots non connectés dans le texte, ce qui devrait augmenter la précision et peut-être le rappel. Par exemple, dans un certificat qui contient sur trois lignes différentes les expressions *insuffisance respiratoire terminale* (J96.9), *insuffisance cardiaque* (I50.9), et *insuffisance rénale chronique sévère* (N18.9), l'application à chaque ligne du classifieur trouve chacun des trois concepts attendus alors que son application globale au certificat entier ne propose pas le troisième, et propose à la place du premier le concept J96.1 (*insuffisance respiratoire chronique*). D'autre part, le classifieur multiétiquette a appris sur des textes entiers à déterminer le nombre de concepts à proposer, et est ainsi biaisé vers un nombre de concepts plus grand que pour un énoncé individuel, ce qui est susceptible d'augmenter le rappel au détriment éventuel de la précision.

Nous insistons sur une condition qui nous semble importante dans la situation étudiée : chaque concept peut être détecté localement, c'est-à-dire à l'intérieur d'un énoncé, comme c'est le cas des entités nommées. En pratique, cette condition n'est pas entièrement satisfaite ici, car le choix d'un concept pour certains diagnostics peut dépendre du contexte. Ainsi, dans un certificat contenant le terme ambigu *choc hémorragique* ainsi que *polytraumatisme*, le classifieur entraîné et appliqué sur les certificats propose bien le concept T79.4 (*choc traumatique*, choc en présence de traumatisme) alors qu'appliqué sur les lignes il propose R57.1 (*choc hypovolémique*), qui est incorrect ici.

Cette étude possède diverses limitations. Ces expériences ont été réalisées sur un seul jeu de données. Leur généralisation à d'autres données reste à tester. Pour des questions de temps, nous n'avons pas

appliqué de validation croisée pour construire les tableaux 3 et 4 : les différences observées sont susceptibles de varier avec des découpages différents en entraînement et test. Notons cependant qu'un certain nombre des différences constatées dans le tableau 4 sont larges.

Les jeux de données employés correspondent aux corpus d'entraînement de CLEF eHealth 2016 et 2017, mais pas à un corpus de test existant. Les résultats présentés ne sont de ce fait pas directement comparables à ceux des participants de la tâche de 2016 ou 2017. On peut noter cependant qu'ils sont du même ordre de grandeur que le résultat du meilleur participant sur le corpus de test de 2016 (F1=84,8 % (Van Mulligen *et al.*, 2016)), qui était évalué au niveau ligne, ou que nos résultats sur ce même corpus de test pour l'union du dictionnaire calibré et du classifieur mono-étiquette (F1=85.9 % (Zweigenbaum & Lavergne, 2016)). Ils sont aussi du même ordre que nos résultats avec le classifieur multi-étiquette sur le corpus de test français de CLEF eHealth 2017 (ligne : F1=86.5 %, certificat : F1=81.7 %, (Zweigenbaum & Lavergne, 2017)), qui surpassent de 5 à 6 points ceux du meilleur participant.

4 Conclusion

Dans le cas des certificats de décès, nous confirmons l'intuition selon laquelle la détection supervisée de concepts au niveau des énoncés (lignes) fonctionne mieux qu'au niveau du texte entier (certificat). Ce constat était bien sûr attendu lorsque l'on dispose d'énoncés annotés ; nous avons observé qu'il s'applique aussi lorsque l'on dispose uniquement de textes annotés, un classifieur entraîné sur les textes entiers donnant en moyenne de meilleurs résultats si on l'applique séparément à chaque énoncé que si on l'applique globalement au texte entier. Nous avons vu également qu'un classifieur multiétiquette était plus robuste mais cependant sensible aussi à cette différence de granularité.

Ces observations ouvrent plusieurs perspectives. Tout d'abord, elles sont directement applicables à des données de nature similaire mais provenant d'une autre source et écrites dans une autre langue : les certificats de décès américains fournis lors de la campagne CLEF eHealth 2017. Contrairement aux certificats français, ces données ne sont annotées qu'au niveau de chaque certificat. Nous avons réalisé des expériences préliminaires qui confirment l'intérêt d'entraîner un classifieur sur ces textes puis de l'appliquer au niveau des lignes individuelles plutôt que des textes entiers. De plus, ces observations incitent à mettre en place des méthodes pour transférer automatiquement au niveau des énoncés les annotations disponibles au niveau des textes. Là aussi, nos premiers essais par annotation automatique d'une partie des énoncés puis ré-entraînement sur le corpus ainsi étendu sont encourageants.

Remerciements

Ce travail a bénéficié d'un soutien de l'action Horizon 2020 Marie Sklodowska-Curie Innovative Training Networks — European Joint doctorate (ITN-EJD) de l'Union européenne, projet No :676207 (MiRoR).

Références

ARONSON A. R. & LANG F.-M. (2010). An overview of MetaMap : historical perspective and

recent advances. *J Am Med Inform Assoc*, **17**(3), 229–36.

LAVERGNE T., NÉVÉOL A., ROBERT A., GROUIN C., REY G. & ZWEIGENBAUM P. (2016). A dataset for ICD-10 coding of death certificates : Creation and usage. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM2016)*, p. 60–69, Osaka, Japan : The COLING 2016 Organizing Committee.

LIN J. & WILBUR W. J. (2007). PubMed related articles : a probabilistic topic-based model for content similarity. *BMC Bioinformatics*, **8**, 423.

NOUVEL D., ERHMANN M. & ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*. Iste Editions.

NÉVÉOL A., COHEN K. B., GROUIN C., HAMON T., LAVERGNE T., KELLY L., GOEURIOT L., REY G., ROBERT A., TANNIER X. & ZWEIGENBAUM P. (2016). Clinical information extraction at the CLEF eHealth evaluation lab 2016. In *CLEF eHealth Evaluation Lab*.

OMS (1993). *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*. Organisation mondiale de la Santé, Genève.

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

TSATSARONIS G., BALIKAS G., MALAKASIOTIS P., PARTALAS I., ZSCHUNKE M., ALVERS M. R., WEISSENBORN D., KRITHARA A., PETRIDIS S., POLYCHRONOPOULOS D., ALMIRANTIS Y., PAVLOPOULOS J., BASKIOTIS N., GALLINARI P., ARTIÈRES T., NGONGA A., HEINO N., ÉRIC GAUSSIER, BARRIO-ALVERS L., SCHROEDER M., ANDROUTSOPOULOS I. & PALIOURAS G. (2015). An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, **16**, 138 :1–138 :28.

VAN MULLIGEN E., AFZAL Z., AKHONDI S. A., VO D. & KORS J. A. (2016). Erasmus MC at CLEF eHealth 2016 : Concept recognition and coding in French texts. In *CLEF 2016 Online Working Notes* : CEUR-WS.

ZHENG J. G., HOWSMON D., ZHANG B., HAHN J., MCGUINNESS D., HENDLER J. & JI H. (2015). Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, **15**(1), S4.

ZWEIGENBAUM P. (2017). Le traitement des langues naturelles dans le contexte de la eSanté. In P. DEGOULET, M. FIESCHI & J. MÉNARD, Eds., *e-santé en perspective*, volume 20 of *Informatique et santé*. Paris : Lavoisier.

ZWEIGENBAUM P. & LAVERGNE T. (2016). Hybrid methods for ICD-10 coding of death certificates. In *Seventh International Workshop on Health Text Mining and Information Analysis*, p. 96–105, Austin, Texas, USA : EMNLP 2016.

ZWEIGENBAUM P. & LAVERGNE T. (2017). Multiple methods for multi-class, multi-label ICD-10 coding of multi-granularity, multilingual death certificates. In *CLEF 2017 Online Working Notes* : CEUR-WS. Sous presse.