Tri Automatique de la Littérature pour les Revues Systématiques

Christopher Norman,[1, 2] Mariska Leeflang,[2] Pierre Zweigenbaum,[1] Aurélie Névéol[1]
(1) LIMSI, CNRS, Université Paris Saclay, 91405 Orsay, France
(2) Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
prenom.nom@limsi.fr, m.m.leeflang@amc.uva.nl

RÉSUMÉ

Les revues systématiques de la littérature dans le domaine biomédical reposent essentiellement sur le travail bibliographique manuel d'experts. Nous évaluons les performances de la classification supervisée pour la découverte automatique d'articles à l'aide de plusieurs définitions des critères d'inclusion. Nous appliquons un modèle de regression logistique sur deux corpus issus de revues systématiques conduites dans le domaine du traitement automatique de la langue et de l'efficacité des médicaments. La classification offre une aire sous la courbe moyenne (AUC) de 0.769 si le classifieur est contruit à partir des jugements experts portés sur les titres et résumés des articles, et de 0.835 si on utilise les jugements portés sur le texte intégral. Ces résultats indiquent l'importance des jugements portés dès le début du processus de sélection pour développer un classifieur efficace pour accélérer l'élaboration des revues systématiques à l'aide d'un algorithme de classification standard.

ABSTRACT

Automatically Ranking the Literature in Support of Systematic Reviews.

Current approaches to document discovery for systematic reviews in biomedicine rely on exhaustive manual screening. We evaluate the performance of classifier based article discovery using different definitions of inclusion criteria. We test a logistic regressor on two datasets created from existing systematic reviews on clinical NLP and drug efficacy, using different criteria to generate positive and negative examples. The classification and ranking achieves an average AUC of 0.769 when relying on gold standard decisions based on title and abstracts of articles, and an AUC of 0.835 when relying on decisions based on full text. Results suggest that inclusion based on title and abstract generalizes to inclusion based on full text, so that references excluded in earlier stages are important for classification, and that common-off-the-shelves algorithms can partially automate the process.

MOTS-CLÉS : Recherche d'Information, Classification Supervisée, Revues Systématiques.

KEYWORDS: Information Retrieval, Supervised Classification, Systematic Reviews.

1 Revues systématiques de la littérature

1.1 Processus de sélection des articles analysés

Les revues systématiques de la littérature ont pour objectif de présenter une synthèse complète et objective de l'ensemble des informations publiées sur une question donnée. Elément essentiel de la médecine factuelle, ces études constituent le niveau de preuve scientifique le plus élevé. Elles contribuent également à l'information du grand public, à l'élaboration de politiques de santé publique.

Cependant, la réalisation d'une revue systématique repose sur des experts et s'avère extrèmement couteuse en temps. Le nombre d'articles à consulter pour effectuer une revue **complète** de la littérature peut se chiffrer à plusieurs milliers, alors que seules quelques dizaines de références passeront le filtre de la méta-analyse finale. Ce processus de sélection, essentiellement manuel, demande plusieurs mois de travail à un collège d'experts.

Le processus de sélection se déroule en plusieurs étapes, comme illustré en figure 1. Un large ensemble d'articles candidats à l'analyse est défini, typiquement grâce à une requête booléenne établie par les experts et soumise à des moteurs recherche spécialisés. Une première sélection est effectuée dans cet ensemble en consultant le titre et le résumé des articles retournés par la requête. Cette étape permet d'éliminer les articles qui ne portent pas sur la question étudiée et les articles qui ne sont pas conformes aux critères d'inclusion de l'étude. Nous désignons les articles éliminés par ce premier filtre comme la catégorie "**N**". Une deuxième sélection est ensuite effectuée sur consultation du texte intégral des articles afin de confirmer que l'ensemble des critères d'inclusion sont bien réunis. Nous désignons les articles éliminés par ce deuxième filtre comme la catégorie "**M**", et les articles finalement retenus comme la catégorie "**Y**".

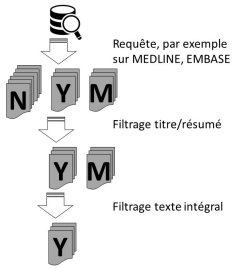

FIGURE 1 – Processus de sélection des articles pour une revue systématique.

Ces deux étapes sont nécessaires, car si un grand nombre d'articles peut être rejetté suite à l'examen du titre et du résumé, la décision finale d'inclusion des articles dans l'analyse ne peut être prise que sur la base du texte intégral car le niveau de détail requis ne se trouve pas dans les résumés. Les décisions d'inclusion réputées difficiles sont prises lors de la deuxième étape, et les articles des catégories Y et M sont listés dans la revue finale, alors que les articles de la catégorie N sont simplement écartés.

Bien que la méthode experte de sélection des articles à inclure dans une revue systématique repose sur plusieurs étapes, les travaux cherchant à automatiser ce processus ont abordé le problème comme une seule étape, celle de la sélection finale des articles à retenir. En effet, le but final de la sélection de documents dans l'élaboration d'une revue systématique est d'identifier les documents satisfaisant les critères d'inclusion. Lors de la construction de jeux de données il n'est donc pas toujours jugé pertinent de distinguer les documents appartenant aux catégories M et N. Nous faisons l'hypothèse que les documents de la catégorie M peuvent être utiles pour entrainer un classifieur et nous étudions différentes configurations des jeux d'entrainement afin définir la configuration la plus efficace.

Dans cet article, nous étudions l'intérêt de distinguer les deux étapes du processus de sélection pour automatiser le tri d'article à inclure dans les revues systématiques. Notre contribution est double : tout d'abord, nous présentons des expériences permettant de choisir une définition adaptée du problème de classification des articles considérés pour inclusion dans les revues systématiques et ensuite, nous nous appuyons sur un corpus existant dans le domaine de l'éfficacité des médicaments et confirmons

les résultats obtenus sur un nouveau corpus dans le domaine orthogonal du TAL.

1.2 Sélection automatique d'articles pour les revues systématiques

De nombreux travaux ont abordé la question de l'automatisation du processus d'élaboration des revues systématiques, avec un succès variable comme l'indique une revue approfondie de la recherche récente sur ce sujet dans le domaine biomédical (O'Mara-Eves *et al.*, 2015). D'autres domaines ont proposé des méthodes pour améliorer la recherche exhaustive de documents, par exemple pour les brevets (Stein *et al.*, 2012) ou les procédures civiles en droit américain (Grossman & Cormack, 2011). Le problème est abordé comme une tâche de classification ou de tri ordonné d'articles. Les méthodes de classification supervisées nécessitent la disponibilité de données d'entrainement, ce qui n'est pas toujours le cas pour des revues dans un nouveau domaine. Cependant, pour des revues de mise à jour ou des revues abordant une nouvelle question dans un domaine déjà exploré, le résultat des revues précédentes peut être exploité.

Ainsi, des classifieurs Bayésiens (Matwin *et al.*, 2010; Matwin & Sazonova, 2012), des SVMs (Support Vector Machines) et des arbres de décision (Bekhuis & Demner-Fushman, 2010) ont permis d'obtenir de bonnes performances. Parmi les méthodes utilisées on dénombre également : des perceptrons (Cohen *et al.*, 2006), kNN (García Adeva *et al.*, 2014), des relations sémantiques (Fiszman *et al.*, 2010), des ontologies (Sun *et al.*, 2012), des modèles linéaires (Shekelle *et al.*, 2012), des classifieurs ensemblistes (Shekelle *et al.*, 2012), l'indexation aléatoire (Jonnalagadda & Petitti, 2014) ou les forêts d'arbres décisionnels (Khabsa *et al.*, 2016). Cependant, peu de ces méthodes ont été évaluées sur des jeux de données communs hormis le corpus Cohen. Nous décrivons ce jeu de données ci-dessous et l'utilisons pour permettre une comparaison directe de nos résultats avec l'état de l'art.

2 Régression logistique pour le tri automatique de documents

2.1 Corpus de Travail

Chaque corpus comporte des références d'articles sous forme de numéro identifiant associé à des métadonnées ainsi qu'à une classification selon les catégories N, M et Y. La distribution des articles issus de chaque étape de sélection est présentée dans le Tableau 1.

Le **corpus du** *Yearbook* est issu de la revue de la littérature en Traitement Automatique de la Langue Clinique effectuée pour le *Yearbook of Medical Informatics* (Névéol & Zweigenbaum, 2016). Ce corpus illustre le cas d'une revue systématique récurrente, proposant des mises à jour successives de l'état de l'art. Pour chaque nouvelle version de la revue, les données issues des revues précédentes peuvent être exploitées pour entraîner un classifieur *intra-topic*.

Le **corpus Cohen**, disponible librement [1], est issu d'un travail précurseur sur la classification automatique d'articles pour les revues systématiques exploitant les données de 15 revues sur l'efficacité des médicaments (Cohen *et al.*, 2006). Ce corpus illustre le cas d'une série de revues sur le même thème général, mais déclinant plusieurs questions. Pour chaque nouvelle question, les données issues des revues précédentes peuvent être exploitées pour entraîner un classifieur *inter-topic*.

1. http://skynet.ohsu.edu/~cohenaa/systematic-drug-class-review-data.html

Topic	Y	M	N	Topic	Y	M	N
Yearbook							
ClinicalNLP (2017)	11	70	244				
ClinicalNLP (2016)	23	60	267				
Cohen							
CalciumChannelBlockers	100	180	938	ProtonPumpInhibitors	51	187	1095
ACEInhibitors	41	142	2361	Triptans	24	194	453
BetaBlockers	42	260	1770	NSAIDS	41	47	305
Opiods	15	33	1867	ADHD	20	64	767
OralHypoglycemics	136	3	364	AtypicalAntipsychotics	146	218	756
Statins	85	88	3292	UrinaryIncontinence	40	38	249
SkeletalMuscleRelaxants	9	25	1609	Estrogens	80	0	288
Antihistamines	16	76	218				

TABLE 1 – Distribution des catégories d'article dans les corpus de travail.

2.2 Méthode de classification

Nous utilisons l'implémentation de regression logistique de sklearn(Pedregosa *et al.*, 2011) entrainé à l'aide d'un gradient stochastique conjugué. Nous avons choisi cette méthode pour sa capacité à traiter rapidement des représentations de données creuses et à effectuer des mises à jour partielles du modèle. Par ailleurs, cette méthode produit un score de confiance associé aux prédictions, ce qui est essentiel pour notre tâche de tri de documents.

Les traits du classifieurs reposent sur les métadonnées extraites de la base MEDLINE pour chacun des articles du corpus : Nous créons des traits avec les sacs-de-n-grammes ($n \leq 3$) issus des titres des résumés, les scores tf-idf ou binaire associés aux mots, les formes racinisées ou originale des mots. Nous considérons séparéments les n-grammes issus des sections "background", "method", "results", et "conclusion" des résumés. Nous considérons également les mots-clés attribués aux articles par les auteurs, les noms des journaux et les types de publication. Pour le corpus Cohen nous extrayons les termes d'indexation MeSH. Pour le corpus Yearbook, ces termes ne sont pas disponibles lors de la préparation de la revue systématique.

Nous abordons deux types de classification : **la classification *intra-topic*** pour laquelle des données d'entrainement et les données de test appartiennent au même thème. Dans le cas du corpus Yearbook, nous entrainons le classifieur sur les données 2016 et nous testons sur les données 2017. Dans le cas du corpus Cohen, nous séparons aléatoirement les données en deux ensembles. Cela présete l'inconvénient de produire des corpus d'entrainement de petite taille, en particulier pour certains thèmes. Afin de pallier cet inconvénénient, nous abordons également **la classification *inter-topic*** pour laquelle des données d'entrainement et les données de test appartiennent à des thèmes différents, mais proches. C'est le cas pour les différents thèmes du corpus Cohen. Dans ce cas, le classifieur est entrainé sur 14 thèmes et testé sur le 15ème thème.

Nous suivons la méthodologie expérimentale utilisée dans la littérature (Cohen *et al.*, 2006; Khabsa *et al.*, 2016). Ainsi, pour la validation croisée inter-topic nous répétons le processus de classification 10 fois. Pour la validation croisée intra-topic nous répétons le processus de classification 5 fois, sur 2 plis. Nous rapportons les résultats moyens (+/- SD) pour chaque expérience en termes d'aire moyenne sous la courbe (AUC) et de score WSS@95 ("work saved over sampling at 95% recall"). Le WSS@95

représente la quantité de travail économisée par un expert qui parcourerait la liste des articles triés par le classifieur jusqu'au seuil de 95% de rappel pour les articles à inclure dans la revue (catégorie Y).

Nous utilisons les paramètres par défaut, sauf dans les cas indiqués ci-dessous. Nous augmentons le poids des exemples positifs en leur affectant un coefficient de 80 afin d'amplifier le coût des erreurs de classification pour les articles pertinents. Nous utilisons la valeur $\alpha = 10^{-4}$ pour la regularisation sur le corpus Cohen, et $\alpha = 0.05$ pour le corpus du Yearbook. Ces valeurs ont été sélectionnées sur la base du résultat d'expériences préliminaires sur le thème (CalciumChannelBlockers) pour le corpus Cohen, et sur la première itération pour le corpus du Yearbook 2016.

3 Résultats

La table 2 présente une comparaison de nos résultats avec l'état de l'art afin de valider la pertinence de notre méthode. Notre classifieur offre des performances supérieures à l'état de l'art pour la mesure WSS@95, et légèrement en dessous pour l'AUC. Notre implémentation semble notament faible lorsque peu d'articles de la catégorie M sont présents (OralHypoGlycemics, Estrogens). On constate cependant qu'aucune méthode ne semble s'imposer sur l'ensemble des thèmes.

	Intertopic			Intratopic			
Mesure / Thème	WSS@95	AUC		WSS@95		AUC	
	NC	NC	(Cohen)	NC	(Khabsa)	NC	(Khabsa)
CalciumChannelBlockers	.13	**.76**	.71	**.40**	.29 (RF)	.83	**.87** (SVM)
ACEInhibitors	.57	**.82**	.81	**.63**	.52 (CNB)	.92	**.95** (RF)
BetaBlockers	.40	**.84**	.80	**.51**	.37 (CNB)	.86	**.89** (RF)
Opiods	.30	**.89**	.86	**.59**	.55 (CNB)	.91	**.91** (RF)
OralHypoglycemics	.07	**.66**	.57	**.11**	.08 (CNB)	.57	**.78** (SVM)
Statins	.27	**.83**	.77	**.44**	.40 (RF)	.87	**.92** (RF)
SkeletalMuscleRelaxants	.24	.83	**.84**	**.43**	.37 (RF)	.74	**.79** (RF)
Antihistamines	.07	**.65**	.62	**.15**	.15 (CNB)	.65	**.72** (SVM)
ProtonPumpInhibitors	.38	**.82**	.79	**.31**	.29 (RF)	.83	**.88** (RF)
Triptans	.46	.82	**.82**	.30	**.31** (RF)	.79	**.91** (SVM)
NSAIDS	.671	**.91**	.90	**.54**	.53 (CNB)	.86	**.95** (SVM)
ADHD	.13	**.59**	.47	.62	**.67** (VP)	.91	**.95** (RF)
AtypicalAntipsychotics	.16	**.76**	.65	**.21**	.21 (CNB)	.78	**.84** (RF)
UrinaryIncontinence	.37	**.89**	.85	**.42**	.41 (RF)	.78	**.89** (SVM)
Estrogens	.18	**.69**	.59	.29	**.38** (CNB)	.69	**.89** (SVM)

TABLE 2 – Comparaison de notre classifieur (NC) sur le corpus Cohen avec (Cohen, 2008) et (Khabsa *et al.*, 2016) pour la classification (Y||MN).

Les Tables 3 et 4 présentent les resultats de nos expériences utilisant différentes combinaisons des classes Y, M et N. Nous indiquons entre parenthèses les classes considérées : par exemple, (Y||MN) dénote l'utilisation de la catégorie Y comme classe positive, les catégories M et N étant considérées comme classe négative. Dans l'expérience (Y|M|N) les catégories Y et M sont utilisées comme classe positive à l'entrainement mais seule la catégorie Y est considérés comme classe positive lors du test.

	(Y||MN)				(YM||N)				(Y|M|N)			
	WSS@95		AUC		WSS@95		AUC		WSS@95		AUC	
Topic	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
ClinicalNLP	.00	.00	.63	.01	.23	.01	.80	.00	.19	.01	.81	.00
CChannelBlockers	.40	.10	.83	.02	.22	.06	.76	.03	.34	.07	.79	.01
ACEInhibitors	.63	.16	.92	.02	.28	.05	.80	.02	.60	.13	.88	.03
BetaBlockers	.51	.16	.86	.03	.19	.05	.73	.03	.48	.21	.83	.02
Opiods	.59	.19	.91	.05	.37	.10	.82	.03	.71	.06	.88	.04
OHypoglycemics	.11	.05	.57	.03	.14	.07	.58	.04	.09	.02	.58	.03
Statins	.44	.18	.87	.02	.25	.09	.78	.03	.42	.10	.86	.02
SkMuscleRelaxants	.43	.22	.74	.11	.26	.18	.83	.06	.45	.12	.75	.06
Antihistamines	.15	.09	.65	.09	.13	.04	.57	.03	.24	.09	.60	.01
PPumpInhibitors	.31	.19	.83	.04	.17	.04	.73	.02	.38	.06	.77	.04
Triptans	.30	.24	.79	.08	.30	.04	.75	.03	.41	.07	.69	.03
NSAIDS	.54	.18	.86	.02	.40	.07	.76	.04	.46	.06	.73	.02
ADHD	.62	.15	.91	.03	.70	.10	.91	.02	.83	.06	.91	.01
AAntipsychotics	.21	.04	.78	.01	.12	.02	.71	.03	.28	.06	.80	.02
UIncontinence	.42	.14	.78	.03	.21	.09	.66	.04	.48	.07	.75	.04
Estrogens	.29	.09	.69	.03	.27	.09	.72	.04	.32	.06	.69	.03

TABLE 3 – Classification Intra-topic à trois classes. Les résultats présentent la moyenne (avg) et l'éccart type (std) obtenus sur 10 itérations (5×2 validation croisée) pour différentes compositions des jeux d'entrainement.

	(Y||M)				(Y||N)				(M||N)			
	WSS@95		AUC		WSS@95		AUC		WSS@95		AUC	
Topic	avg	std	avg	std	avg	std	avg	std	avg	std	avg	std
ClinicalNLP	.01	.00	.48	.01	.02	.00	.74	.00	.26	.00	.79	.00
CChannelBlockers	.14	.04	.59	.03	.42	.11	.85	.02	.21	.07	.74	.03
ACEInhibitors	.17	.08	.63	.06	.41	.37	.92	.03	.26	.06	.77	.02
BetaBlockers	.38	.10	.74	.02	.52	.14	.87	.03	.19	.03	.71	.02
Opiods	.13	.10	.53	.01	.59	.21	.91	.06	.25	.18	.76	.05
OHypoglycemics	.06	.00	.39	.17	.11	.04	.58	.03	.75	.19	.83	.11
Statins	.13	.05	.56	.04	.44	.18	.88	.05	.24	.09	.71	.03
SkMuscleRelaxants	.24	.14	.55	.02	.30	.15	.67	.08	.23	.16	.80	.07
Antihistamines	.20	.17	.55	.06	.16	.09	.70	.03	.13	.07	.58	.04
PPumpInhibitors	.16	.05	.58	.02	.42	.17	.85	.03	.12	.05	.69	.03
Triptans	.20	.13	.70	.07	.44	.24	.88	.04	.27	.06	.75	.03
NSAIDS	.13	.05	.58	.06	.48	.19	.85	.02	.32	.09	.72	.03
ADHD	.19	.14	.59	.09	.71	.17	.94	.02	.64	.17	.92	.01
AAntipsychotics	.11	.02	.55	.02	.26	.11	.79	.03	.11	03	.63	.03
UIncontinence	.09	.04	.55	.02	.43	.16	.79	.03	.12	.10	.59	.05
Estrogens	-	-	-	-	.23	.03	.69	.04	-	-	-	-

TABLE 4 – Classification Intratopic à deux classes. Les résultats présentent la moyenne (avg) et l'éccart type (std) obtenus sur 10 itérations (5×2 validation croisée) pour différentes compositions des jeux d'entrainement.

Globalement, nos résultats montrent l'intérêt de disposer des documents de la catégorie N pour entraîner nos classifieurs. En effet, les performances les plus médiocres sont obtenues avec la catégorisation (Y||M), qui correspond pourtant à la deuxième étape de filtrage, considérée plus importante et mieux documentée.

Dans la Table 3, on observe des résultats similaires pour les classifications (Y|M|N) et (YM||N), ce qui indique que les exemples de la catégorie M peuvent être utilement utilisés comme substituts aux exemples de la catégorie Y lorsque ceux-ci sont rares. Par ailleurs, les résultats de la classification (Y||MN) sont largement inférieurs, ce qui confirme notre hypothèse et montre l'importance de bien distinguer la catégorie M de la catégorie N.

4 Conclusion et perspectives

En accord avec la littérature, nos résultats indiquent qu'un classifieur entrainé sur le titre et le résumé d'articles offre de bonnes performances pour le tri d'articles sur des critères d'inclusion majoritairement présent dans le texte intégral. Cependant, pour atteindre ce résultat, il est important de disposer d'exemples d'articles exclus sur la base du résumé seul.

Ainsi, pour construire un corpus de référence qui permette d'évaluer ou d'entrainer des outils automatiques de tri de la littérature pour les revues systématiques, il est nécessaire de disposer d'une catégorisation des articles qui reflète les deux étapes du processus manuel de filtrage.

Dans la suite de ce travail, nous envisageons plusieurs directions : tout d'abord, nous souhaitons évaluer la généralisabilité de la classification aux revues systématiques des études sur l'évaluation des tests diagnostiques, réputées fastidieuses de par la difficulté de construire des requêtes permettant de réduire l'ensemble des documents à considérer. Ensuite, nous prévoyons d'étudier l'apport du retour de pertinence (*relevance feedback*) en particulier pour les thèmes pour lesquels il n'est pas possible de construire d'emblée un classifieur intra-topic. L'intégration du retour de pertinence correspondrait en outre à un cas pratique d'utilisation de méthodes de classification automatique pour la sélection d'articles en vue d'une revue systématique où les experts valideraient la classification des articles dans l'ordre proposé par l'outil - produisant ainsi des jugements intégrables dynamiquement dans le classifieur.

Remerciements

Ce travail a bénéficié d'un financement du programme européen Horizon 2020 Marie Sklodowska Curie Innovative Training Networks—European Joint doctorate (ITN-EJD) No :676207, Methods in Research on Research (MiRoR).

Références

BEKHUIS T. & DEMNER-FUSHMAN D. (2010). Towards automating the initial screening phase of a systematic review. *Studies in Health Technology and Informatics*, **160**(PART 1), 146–150.

COHEN A. M. (2008). Optimizing feature representation for automated systematic review work prioritization. *AMIA Annual Symposium proceedings*, p. 121–5.

COHEN A. M., HERSH W. R., PETERSON K. & YEN P. (2006). Reducing Workload in Systematic Review Preparation Using Automated Citation Classification. p. 206–219.

FISZMAN M., BRAY B., SHIN D., KILICOGLU H., BENNETT G., BODENREIDER O. & RINDFLESCH T. (2010). Combining Relevance Assignment with Quality of the Evidence to Support Guideline Development. *Stud Health Technol Inform*, **160**(1), 709—-713.

GARCÍA ADEVA J. J., PIKATZA ATXA J. M., UBEDA CARRILLO M. & ANSUATEGI ZENGOTITABENGOA E. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, **41**(4 PART 1), 1498–1508.

GROSSMAN M. & CORMACK G. (2011). Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Richmond Journal of Law and Technology*, **17**(3).

JONNALAGADDA S. R. & PETITTI D. (2014). A new iterative method to reduce workload in the systematic review process. *Int J Comput Biol Drug Des*, **6**(0), 5–17.

KHABSA M., ELMAGARMID A., ILYAS I., HAMMADY H. & OUZZANI M. (2016). Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, **102**(3), 465–482.

MATWIN S., KOUZNETSOV A., INKPEN D., FRUNZA O. & O'BLENIS P. (2010). A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association : JAMIA*, **17**(4), 446–53.

MATWIN S. & SAZONOVA V. (2012). Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, **19**(5), 917–917.

NÉVÉOL A. & ZWEIGENBAUM P. (2016). Clinical natural language processing in 2015 : Leveraging the variety of texts of clinical interest. *IMIA Yearbook*, p. 234–239.

O'MARA-EVES A., THOMAS J., MCNAUGHT J., MIWA M. & ANANIADOU S. (2015). Using text mining for study identification in systematic reviews : a systematic review of current approaches. *Systematic reviews*, **4**(1), 5.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O. *et al.* (2011). Scikit-learn : Machine learning in python. *Journal of Machine Learning Research*, **12**(Oct), 2825–2830.

SHEKELLE P. G., DALAL S. R. & SHETTY K. D. (2012). A Pilot Study Using Machine Learning and Domain Knowledge To Facilitate Comparative Effectiveness Review Updating. *AHRQ*.

STEIN B., HOPPE D. & GOLLUB T. (2012). The impact of spelling errors on patent search. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, p. 570–579 : Association for Computational Linguistics.

SUN Y. B., YANG Y., ZHANG H., ZHANG W. & WANG Q. (2012). Towards evidence-based ontology for supporting systematic literature review. *Evaluation and Assessment in Software Engineering*, **2012**(1), 171–175.