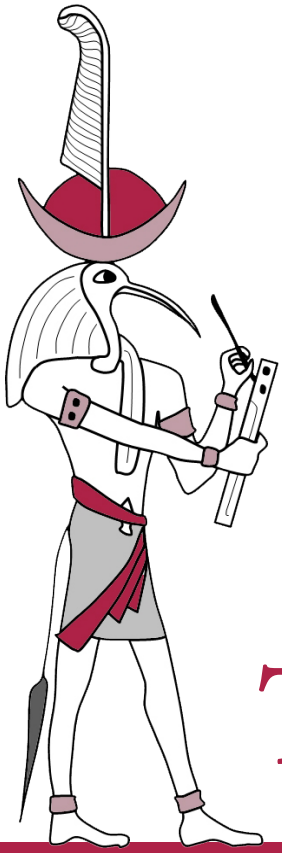
Terminologica



TOTh 2019

Terminologie & Ontologie : Théories et Applications

Presses Universitaires Savoie Mont Blanc

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2019

Le Bourget du Lac – 6 & 7 juin 2019

Les ouvrages TOTh précédents sont disponibles sur le site du Comptoir des Presses d'Universités (www.lcdpu.fr) ou auprès de : contact@toth.condillac.org

Éditeur : Presses Universitaires Savoie Mont Blanc
27 rue Marcoz
BP 1104
73011 CHAMBÉRY CEDEX
www.univ-smb.fr

Réalisation : C. Brun, C. Roche
Collection « Terminologica »
ISBN : 978-2-919732-80-7
ISSN : 2607-5008
Dépôt légal : juillet 2020

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2019

Le Bourget du Lac – 6 & 7 juin 2019

http://toth.condillac.org

avec le soutien de :
Université Savoie Mont Blanc
École d'ingénieurs Polytech Annecy Chambéry

Presses Universitaires Savoie Mont Blanc
Collection « Terminologica »

Comité scientifique

Président du Comité scientifique : Christophe Roche

Comité de pilotage

Rute Costa	Universidade Nova de Lisboa
Humbley John	Université Paris 7
Kockaert Hendrik	University of Leuven
Christophe Roche	Université Savoie Mont Blanc

Comité de programme 2019

Le comité de programme est constitué chaque année à partir du comité scientifique de TOTh en fonction des soumissions reçues. La composition du comité scientifique est accessible à l'adresse suivante : http://toth.condillac.org/committees

Guadelupe Aguado	Universidad Politécnica de Madrid – Spain
Amparo Alcina	Universitat Jaume I – Spain
Bruno Bachimont	Université Technologie de Compiègne – France
Jean-Paul Barthès	Université Technologie de Compiègne – France
Christopher Brewster	TNO – The Netherlands
Danielle Candel	CNRS, Université Paris Diderot – France
Sylviane Cardey	Université de Franche-Comté – France
Stéphane Chaudiron	Université de Lille 3 – France
Manuel Célio Conceição	Universidade do Algarve – Portugal
Rute Costa	Universidade NOVA de Lisboa – Portugal
Bruno Courbon	Université Laval – Canada
Lyne Da Sylva	Université de Montréal – Canada
Luc Damas	Université Savoie Mont-Blanc – France
Éric De La Clergery	INRIA – France
Dardo De Vecchi	Kedge Business School – France
Valérie Delavigne	Université Paris 3 – France
Sylvie Després	Université Paris 13 – France
Juan Carlos Diaz Vasquez	EAFIT University – Colombia
Hanne Erdman Thomsen	Copenhagen Business School – Denmark
Pamela Faber	Universidad de Granada – Spain
Christiane Fellbaum	Princeton University – USA
Iolanda Galanes	Universidade de Vigo – Spain
Christian Galinski	INFOTERM – Austria
François Gaudin	Université de Rouen – France
Teodora Ghiviriga	Alexandru Ioan Cuza University – Romania
Jean-Yves Gresser	ancien Directeur à la Banque de France – France
Ollivier Haemmerlé	Université de Toulouse – France
Gernot Hebenstreit	University of Graz – Austria
Amanda Hicks	University of Florida – USA
John Humbley	Université Paris 7 – France

Avant-propos

La Terminologie est une discipline scientifique à part entière qui puise à de nombreux domaines dont la linguistique, la théorie de la connaissance et la logique. Pour que cette diversité soit une richesse, il faut lui offrir un cadre approprié au sein duquel elle puisse s'exprimer et s'épanouir : c'est une des raisons d'être des Conférences TOTh créées en 2007. A ces conférences « mères » qui se tiennent chaque année à l'Université Savoie Mont-Blanc sont associées depuis 2011 les Journées d'étude TOTh dédiées à un thème plus spécifique organisées par une institution partenaire.

Dans ce contexte, la formation et la transmission des connaissances jouent un rôle essentiel. La *Formation TOTh* précédant la Conférence se déroule sur deux années consécutives dédiées pour l'une à la dimension linguistique et pour l'autre à la dimension conceptuelle de la terminologie, deux dimensions étroitement liées.

A la présentation de travaux sélectionnés par un Comité de programme international, la *Conférence TOTh* inclut une *Conférence invitée* et, selon les années, une *Disputatio*. La première, donnée par une personnalité reconnue dans son domaine vise l'ouverture à d'autres approches de la langue et de la connaissance. La seconde, à travers une lecture commentée effectuée par un membre du comité scientifique, renoue avec une forme d'enseignement et de recherche héritée de la scolastique.

Christian Galinski de Infoterm, a ouvert la conférence sur le sujet de « *The emergence of terminology science and terminological activities* ».

Cette année, comme en 2018, nous n'avons pas inclus de *Disputatio* par manque de temps. En effet, pour la première fois, TOTh a accueilli une session satellite, en parallèle avec la conférence, sur le thème de « Terminology and Text Mining » en lien direct avec les thèmes de TOTh. Nous avons également dédié une session de la conférence au projet Européen ELEXIS.

Les 29 communications et les 3 posters ont permis d'aborder de nombreux sujets tant théoriques que pratiques, autant d'exemples de la diversité et de la richesse de notre discipline. Je vous invite à découvrir à travers ces actes les 24 interventions qui ont donné lieu à publication.

Avant de vous souhaiter bonne lecture, j'aimerais terminer en remerciant tous les participants pour la richesse des débats et des moments partagés.

Christophe Roche
Président du comité scientifique

Sommaire

La technicité des termes : le *v-tech* comme paramètre d'évaluation

Federica Vezzani

Département d'études linguistiques et littéraires (DiSLL)
Piazzetta Gianfranco Folena, 1, 35137, Padoue - Italie
Université de Padoue
federica.vezzani@phd.unipd.it
http://www.dei.unipd.it/~vezzanif/

Résumé. Dans cette étude, nous proposons une perspective neuve du concept de poids des termes techniques en nous concentrant sur la notion de « technicité » comme propriété sémantique de l'unité linguistique elle-même. L'idée de base est que la valeur de technicité d'un terme est inversement proportionnelle à sa nature polysémique. Nous formalisons la formule *v-tech* et effectuons une évaluation expérimentale afin de 1) comparer la valeur *v-tech* avec d'autres mesures de *termhood* (termicité ou termitude) généralement calculées sur la fréquence d'occurrence des termes dans les collections, et 2) intégrer la formule *v-tech* dans le score d'un modèle de récupération de documents pertinents pour un travail de revue systématique dans le domaine médical.

1. Introduction

Cette étude s'inscrit dans le contexte de la terminologie computationnelle comme domaine d'étude récent qui vise à adopter des méthodes computationnelles et quantitatives afin de mener des recherches terminologiques et qualitatives (Bourigault *et al*. 2001, Foo 2012, Drouin *et al*. 2018). Dans la littérature, de nombreuses études sont principalement axées sur l'extraction automatique de termes à partir d'un corpus de documents spécialisés au moyen d'approches 1) linguistiques, 2) statistiques et 2) hybrides (Nakagawa 2001, Vu *et al*. 2008, Amjadian *et al*. 2018, Simon et Kešelj 2018, Sandoval *et al*. 2018). L'acquisition de termes liés et pertinents à un domaine spécifique de l'activité humaine est effectuée automatiquement à l'aide d'approches quantitatives importées du domaine de la recherche d'information : *Term*

frequency-Inverse Document Frequency (TF-IDF) (Salton et Yang 1973), *Mutual Information* (Church et Hanks 1990), T-Score (Church *et al*. 1991), C/NC (Frantzi *et al*. 1998).[1] En outre, des ressources spécifiquement conçues pour cette tâche ont été élaborées afin d'augmenter les performances d'extraction : TermoStat (Drouin 2003), BiTermEx (Planas 2012), TermEvaluator (Inkpen *et al*. 2016) et le récent projet MultiMedica (Sandoval *et al*. 2018) pour l'acquisition des termes concernant le domaine médical. L'importance d'une résolution efficace de cette tâche se reflète enfin dans de nombreux domaines de recherche. L'extraction automatique de termes permet d'effectuer des tâches liées à la recherche d'information (comme le repérage de documents pertinents pour une requête donnée), à la fouille de textes (*text mining*), à la construction de ressources terminologiques, etc.

Le point de départ de tous ces travaux concerne l'identification des termes candidats et, par conséquent, le filtrage entre, d'une part, les mots d'ordre général et d'autre part, les termes spécifiques d'un domaine donné. En effet, toutes les études précédemment citées portent (plus ou moins explicitement) sur le concept de « poids » des termes dans une collection des documents afin d'indiquer les différents degrés de pertinence à un domaine. Ce concept a été exprimé, au fil du temps, à travers différentes dénominations. En 1972, Karen Spärck Jones (Sparck Jones 1972) définissait la notion de « spécificité » des termes, comme une valeur calculable en fonction de la fréquence d'apparition des termes dans une collection des documents :

> "[…] the specificity of an individual term is the level of detail at which a given concept is represented."

> "[…] terms should be weighted according to collection frequency, so that matches on less frequent, more specific, terms are of greater value than matches on frequent terms."

En 1996, Kageura et Umino (Kageura et Umino 1996) introduisaient le concept de « termhood » [termicité ou termitude (Humbley 2016)] afin d'indiquer le degré de relation d'une unité linguistique à des concepts spécifiques pour un domaine, une valeur qui peut également être calculée à l'aide d'approches statistiques. Dans ce sens, le degré de termicité d'un terme est donc une valeur déterminante pour la tâche d'extraction automatique des termes et

1 Pour une synthèse de principaux critères, voir le tableau de (Roche 2018).

repose, en général, sur la fréquence d'apparition d'un terme candidat dans le corpus analysé.

Dans ce contexte, nous proposons une perspective différente du concept de « poids » d'un terme technique en nous concentrant sur la notion de « technicité » comme propriété sémantique intrinsèque au terme lui-même. Après sa description théorique, nous procédons à la formalisation de ce concept au moyen d'une fonction qui pondère le degré d'association d'un terme avec un domaine d'intérêt spécifique. En ce sens, notre objectif n'est pas de fournir une nouvelle méthode pour l'extraction automatique des termes, mais plutôt de définir un nouveau paramètre pour leur évaluation.

Cet article est donc organisé comme suit : dans la section 2, nous définissons le concept de « technicité » et nous formalisons cette propriété au moyen de la formule *v-tech*. Dans la section 3, nous présentons une analyse expérimentale menée afin i) de calculer à la fois le degré de technicité et la valeur de termicité de termes médicaux en langue anglaise et ii) d'évaluer la mesure *v-tech* pour la tâche de repérage de documents, en particulier, pour les revues systématiques dans le domaine médical. Enfin, dans la section 4, nous tirons nos conclusions et décrivons nos perspectives.

2. La valeur de technicité d'un terme : le *v-tech*

La raison de ce travail découle de l'analyse des méthodes actuelles pour l'extraction automatique de termes et de la définition implicite de poids d'un terme comme valeur qui dépend de la fréquence et de la distribution de ses occurrences dans un document et/ou un ensemble de documents. Pour cette raison, ces méthodes sont strictement dépendantes du corpus, et le poids d'un terme peut varier en fonction de la collection analysée.

Dans cette étude, nous proposons plutôt de considérer le poids d'un terme comme une valeur qui correspond au niveau de technicité d'un terme pour un domaine donné. La « technicité » est donc une propriété intrinsèque au terme lui-même et pas une valeur statistique dépendant du corpus. Tout en définissant cette propriété, nous ne faisons pas référence aux concepts précédents de 1) « spécificité », comme le niveau de détail auquel un concept donné est représenté, ou 2) « termicité », comme la propriété d'être ou non un terme. Nous ne faisons pas non plus référence à la technicité dans sa connotation négative de 3) « difficulté » de compréhension : à cet égard, de nombreuses études se concentrent, par exemple, sur la terminologie utilisée dans le dialogue patient-médecin et sur les problèmes connexes en matière de com-

préhension et de lisibilité (Tran *et al*. 2009, Bouamor *et al*. 2018, Grabar et Hamon 2016, Ley 1988, Vecchiato et Gerolimich 2013).

La valeur de « technicité » que nous proposons ici dépend du degré d'association du terme à un domaine d'intérêt. Pour fournir une explication intuitive, nous considérons les termes « quadrantectomie » et « patient » qui relèvent du domaine médical. Les deux termes semblent avoir un poids et un degré de technicité différent : le terme « quadrantectomie »[2] désigne un concept unique dans le domaine de la chirurgie et il est répandu chez les spécialistes de ce seul domaine, alors que le terme « patient »[3] véhicule des significations différentes dans plusieurs domaines (médecine, philosophie et linguistique) et son usage est répandu dans un plus grand nombre de domaines de spécialité, et même dans le langage courant. Si nous excluons de cette analyse la diffusion, en tant que concept proche de la fréquence d'apparition d'un terme dans un corpus réel ou imaginaire, nous nous concentrons sur la définition de « technicité » comme propriété dont la valeur est inversement proportionnelle au nombre de domaines dans lesquels un terme apparaît. En ce sens, plus un terme est polysémique, c'est-à-dire plus il relève de plusieurs domaines, moins il sera technique, car il ne sera ni monoréférentiel ni exclusif. Une caractéristique souhaitable de la terminologie employée pour un domaine est celle d'être monoréférentielle et spécifique (Guilbert 1973) : la relation entre signe et référent devrait être univoque afin d'éviter les ambigüités et la polyvalence du point de vue sémantique. Or, un signe linguistique (un terme) qui désigne plus de référents perd en « technicité », car il n'est plus exclusif d'un seul domaine. À partir de cette définition, nous avons formalisé ce principe en attribuant une valeur numérique à la technicité d'un terme. Nous introduisons donc la valeur que nous appellerons *v-tech* par la formule suivante :

$$\text{v-tech (t)} = \begin{cases} e^{-\lambda d_t}, & d_t > 0 \\ 0, & d_t = ? \end{cases}$$

où *t* est le terme pour lequel la valeur de technicité (*v-tech*) est calculée, où d_t est le nombre de domaines dans lesquels un terme *t* apparaît, et où l est un paramètre qui adapte la rapidité avec laquelle le terme *t* perd en technicité a1 mesure qu'augmente le nombre de domaines qui l'adoptent. Le principe à la base de *v-tech* est que les termes polysémiques, ayant plusieurs domaines d'utilisation, auront une valeur qui tendra vers 0. Les termes n'ayant pas un

2 http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=26527442
3 http://www.cnrtl.fr/definition/patient

domaine clairement explicité dans une ressource auront une valeur de *v-tech* égale à 0.

L'image suivante (FIG. 1) illustre l'évolution de *v-tech* en fonction de d_t lorsque le paramètre l varie :

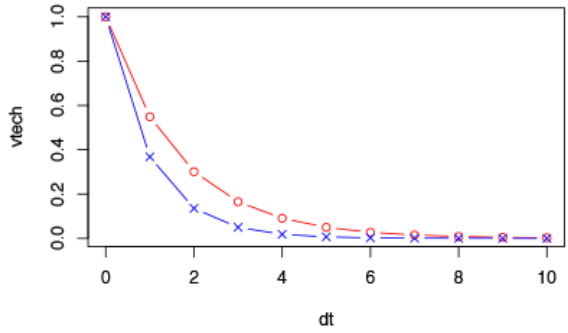


FIG. 1 – *Valeurs de v-tech pour l=0,6 (ligne rouge) et l=1 (ligne bleu)*

Or, pour calculer le score *v-tech*, il est nécessaire de s'appuyer sur des ressources linguistiques montrant tous les domaines associés à une unité linguistique. Par conséquent, le poids d'un terme n'est plus une valeur qui dépend du corpus et calculée en fonction de la fréquence de ses occurrences, mais il devient une valeur dépendant des ressources et basée sur l'exhaustivité des données relatives aux domaines dans lesquels le terme est utilisé.

À notre connaissance, BabelNet est actuellement la ressource la plus complète et la plus structurée qui contient les domaines associés aux termes (Camacho-Collados et Navigli 2017). Cependant, il y a d'autres ressources qui, pour certains termes techniques, collectent plus de domaines que BabelNet. Pour cette raison, afin de calculer la valeur *v-tech* de manière complète et précise, nous avons mené les expériences décrites dans la section suivante en rassemblant les informations fournies par BabelNet, Termium Plus,[4] la base

4 https://www.btb.termiumplus.gc.ca

de données IATE,[5] le dictionnaire en ligne Merriam-Webster[6] et le Grand Dictionnaire Terminologique.[7]

3. Expériences

Dans cette section, nous présentons une analyse expérimentale menée dans le but de : 1) donner un « poids » à un ensemble de termes en comparant la valeur de *v-tech*, telle que définie ci-dessus, à d'autres mesures de termicité basées sur la collection, 2) évaluer l'intégration de la formule *v-tech* dans le score d'un modèle de repérage de documents pour la tâche de revue systématique. Une revue systématique consiste à identifier et à collecter toutes les études, publiées ou non, traitant d'un sujet donné.[8]

Pour cette analyse, nous avons choisi le domaine médical et nous avons utilisé la collection de documents en langue anglaise fournis par le colloque CLEF 2018 (*Conference and Labs of the Evaluation Forum*),[9] pour l'accomplissement de la tâche nommée « *e-Health Technology Assisted Reviews in Empirical Medicine* » (Kanoulas *et al*. 2018). L'ensemble de données comprend : i) 30 sujets médicaux, à savoir les besoins d'informations médicales à satisfaire fournis par un médecin lors de l'accomplissement de revues systématiques ; ii) un ensemble de documents de PubMed[10] ; et iii) un ensemble de jugements de pertinence.

3.1. Le paramètre *V-tech* confronté à d'autres mesures de termicité

Dans la première partie de notre analyse, nous nous concentrons sur l'attribution d'un score représentant le poids d'un ensemble de termes médicaux. Notre but est de comparer la formule *v-tech* avec d'autres mesures de termicité basées sur le corpus. À cet égard, nous procédons d'abord à l'extraction manuelle des termes médicaux identifiés dans l'ensemble des 30 sujets de la collection CLEF 2018. Ensuite, pour chaque terme, nous avons effectué une analyse qualitative en vérifiant tous les domaines d'utilisation de ces termes sur les ressources mentionnées ci-dessus. Nous avons collecté un total de 192 termes ayant un comportement sémantique différent : 104 termes

5 https://iate.europa.eu/home
6 https://www.merriam-webster.com
7 http://www.granddictionnaire.com
8 https://ccf.cochrane.org/revues-cochrane
9 http://clef2018.clef-initiative.eu/index.php
10 https://www.ncbi.nlm.nih.gov/pubmed

monosémiques apparaissent dans un seul domaine (médecine) ou sous-domaine (oncologie, pathologie, chirurgie) ; les 88 autres termes polysémiques envisagent jusqu'à 13 domaines d'utilisation. Dans le TAB.1, nous pouvons observer quelques résultats de notre analyse qualitative : des termes comme « cytology » et « radiculopathy », en tant que monosémiques, ont une valeur de *v-tech* élevée, alors que d'autres termes comme « screening » et « marker » qui sont bien attestés dans plusieurs domaines, ont plutôt une valeur de *v-tech* très basse.

Ensuite, nous avons calculé la fréquence d'occurrence TF (*Term Frequency*) de ces termes dans la collection de documents de PubMed, la fréquence de documents DF (*Document Frequency*) dans lesquels apparaissent les termes analysés et le C-Value visant à donner un poids aux termes complexes (par exemple « globule rouge » ou « douleur musculaire »). Dans le TAB. 2, nous montrons un exemple de trois termes pour deux sujets (A et B correspondent respectivement aux sujets CD010680 et CD008892 de la collection CLEF 2018) avec leur TF, DF, C-Value et la valeur *v-tech* calculée avec l=0.6. Si l'on considère ces deux sous-ensembles de documents, les trois termes ont des fréquences variables et la valeur de leur termicité (TF, DF, C-Value) change significativement en fonction des documents analysés. Si l'on confronte les trois mesures de termicité basées sur le corpus, le terme « pulmonary », par exemple, a un poids complètement différent dans les deux collections de documents A et B. D'autre part, le poids de ces termes basés sur la valeur de *v-tech* reste exactement le même, dans la mesure où la technicité est calculée sur les domaines d'utilisation et définie sur la ressource linguistique.

Terme	Domaine	Ressource
Diagnostic	Information Technology, Medicine, Systems Analysis, Meteorological Forecasting, Psychology, Servicing and Maintenance, Law	BabelNet, Termium plus, IATE
Endemic	Biology, Epidemiology	BabelNet, Termium plus
Cytology	Biology	BabelNet, Merriam-Webster

Terme	Domaine	Ressource
Screening	Economics, Security, Air Transport, Water Treatment, Records Managment, Football, Epidemiology, Cinematography, Press, Waste Managment, Law, Medicine	BabelNet, Termium plus
Marker	Aeronautical, agriculture, army, art, biology, electronic, geology, linguistic, medicine, information science, sociology, sport, telecommunication	BabelNet, Le grand dictionnaire terminologique
Radiculopathy	Pathology	BabelNet, Merriam-Webster

Tab. 1 – *Liste partielle de termes médicaux, domaines d'utilisation et ressources employées.*

Topic	Term	TF	DF	C-Value	v-tech
A	Pulmonary	994	303	993	0.549
A	Typhoid	1	1	0	0.549
A	Resonance	392	273	391	0.015
B	Pulmonary	8	5	6.86	0.549
B	Typhoid	29	13	27.5	0.549
B	Resonance	8	6	6.85	0.015

Tab. 2 – *Comparaison entre TF, DF, C-Value et V-tech de trois termes dans deux sujets.*

3.2. Revues systématiques médicales avec *V-tech*

Dans cette deuxième expérience, nous visons à évaluer l'impact de la valeur *v-tech* pour la tâche de revues systématiques dans le domaine médical, c'est-à-dire pour le repérage de tous les documents pertinents à un sujet médical donné. Pour accomplir cette tâche, de nombreux modèles de repérage permettent d'attribuer un poids aux termes qui constituent la requête : voir le BM25 (Robertson et Zaragoza 2009). Notre hypothèse est qu'en combinant le poids donné par le modèle de base et le poids calculé à partir de la fonction

v-tech, la précision des modèles de repérage s'améliore en termes de documents pertinents récupérés.

Comme modèle de base, nous utilisons la variante CAL (*Continuous Active Learning)* (Di Nunzio 2018) du modèle de repérage BM25. En particulier, suivant l'approche suggérée par (Ventura 2014) avec un document *d* et une requête *q* le nouveau score BM25 est :

$$score(d,q) = \sum_{t \in q \cap d} w_t^{BM25}(1 + \text{v-tech(t)})$$

où, pour chaque terme de la requête qui apparaît dans le document, nous multiplions le poids du BM25 du terme w_t^{BM25} par sa valeur *v-tech*. Nous ajoutons 1 à la valeur *v-tech* afin de prendre en compte les termes pour lesquels le nombre de domaines est inconnu et qui ont, par définition, une valeur de *v-tech* égale à 0.

Dans l'image suivante (FIG. 2), nous présentons les résultats obtenus, en termes de repérage sur la moyenne des 30 sujets, de l'application du modèle de base (baseline) et du modèle de base combiné avec la valeur *v-tech* (v-tech l=0.6). Les mesures d'évaluation sont : 1) la précision à *k* de documents extraits (P@k, c'est-à-dire le rapport entre le nombre de documents pertinents dans les premiers *k* documents et la variable *k*), et 2) le rappel à *R* documents pertinents (c'est-à-dire le nombre de documents pertinents repérés divisé par le nombre *R* total de documents pertinents présents dans la collection).

Les meilleurs résultats (en gras) confirment que la combinaison de la valeur *v-tech* dans le score de repérage du BM25 augmente la précision des documents pertinents récupérés dans les premiers 10, 20, et 50 documents. En outre, le rappel global améliore par rapport au modèle de base.

model	# docs	# rel docs	# rel ret	P@10	P@20	P@50	Recall
baseline	217507	3964	886	0.340	0.345	0.348	0.447
vtech $\lambda = 0.6$	217507	3964	**922**	**0.353**	**0.358**	**0.354**	**0.460**

FIG. 2 – *Résultats moyens sur 30 sujets obtenus en utilisant le modèle de base BM25 et le modèle combiné avec le v-tech.*

4. Conclusions et perspectives

Cette étude repense le concept de poids d'un terme en se concentrant sur la notion de « technicité » comme propriété intrinsèque des termes. L'idée de base est que la valeur de technicité d'un terme est inversement proportionnelle à sa nature polysémique. Sur la base de ce principe, nous avons développé une définition formelle à travers la formule *v-tech*. Dans les expériences menées, nous avons montré que le poids donné par le *v-tech* est une valeur différente d'autres mesures statistiques basées sur un corpus (telles que TF-IDF, C-value, etc.), puisqu'elle est calculée à partir des informations fournies par les ressources linguistiques sur les domaines d'utilisation. De plus, l'intégration de la valeur *v-tech* dans le modèle de base du BM25 a montré que les performances de repérage s'améliorent en moyenne dans des tâches spécifiques telles que la récupération de documents médicaux pour les revues systématiques.

À partir de cette première analyse, nous avons constaté la nécessité de créer une ressource terminologique structurée qui puisse intégrer complètement les informations en ligne. Dans les expériences menées jusqu'à présent, nous avons rassemblé les informations fournies par des ressources différentes. En outre, nous nous proposons de travailler sur un corpus de textes en langue française afin de comparer le différent degré de technicité des termes, telle que celle-ci a été définie, dans une perspective multilingue. Enfin, spécifiquement pour le domaine médical, nous visons à approfondir la relation entre technicité et difficulté de compréhension dans le contexte du dialogue médecin-patient.

Remerciements

Je tiens à remercier le Professeur Giorgio Maria Di Nunzio du Département d'Ingénierie de l'Information (Université de Padoue) pour son aide dans la définition formelle de *v-tech* et pour les expériences menées jusqu'à présent.

Références

Amjadian, Ehsan, Diana Inkpen, T Sima Paribakht, and Farahnaz Faez. 2018. "Distributed Specificity for Automatic Terminology Extraction." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 24 (1): 23-40.

Bouamor, Dhouha, Leonardo Campillos Llanos, Anne-Laure Ligozat, Sophie Rosset, and Pierre Zweigenbaum. 2016. "Transfer-Based Learning-to-Rank Assessment of Medical Term Technicality." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2312-2316.

Bourigault, Didier, Christian Jacquemin, and Marie-Claude L'Homme. 2001. *Recent Advances in Computational Terminology*. Vol. 2. John Benjamins Publishing.

Camacho-Collados, Jose, and Roberto Navigli. 2017. "BabelDomains : Large-Scale Domain Labeling of Lexical Resources." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 2, Short Papers*, 223-228.

Church, Kenneth, William Gale, Patrick Hanks, and Donald Hindle. 1991. "Using Statistics in Lexical Analysis." *Lexical Acquisition : Exploiting on-Line Resources to Build a Lexicon* 115 : 164.

Church, Kenneth Ward, and Patrick Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1) : 22-29.

Di Nunzio, Giorgio Maria. 2018. "A Study of an Automatic Stopping Strategy for Technologically Assisted Medical Reviews." In *European Conference on Information Retrieval*, 672-677. Springer.

Drouin, Patrick. 2003. "Term Extraction Using Non-Technical Corpora as a Point of Leverage." *Terminology* 9 (1) : 99-115.

Drouin, Patrick, Natalia Grabar, Thierry Hamon, Kyo Kageura, and Koichi Takeuchi. 2018. "Computational Terminology and Filtering of Terminological Information." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 24 (1) : 1-6.

Foo, Jody. 2012. "Computational Terminology : Exploring Bilingual and Monolingual Term Extraction." PhD Thesis, Linköping University Electronic Press.

Frantzi, Katerina T, Sophia Ananiadou, and Junichi Tsujii. 1998. "The C-Value/Nc-Value Method of Automatic Recognition for Multi-Word Terms." In *International Conference on Theory and Practice of Digital Libraries*, 585-604. Springer.

Grabar, Natalia, and Thierry Hamon. 2016. "A Large Rated Lexicon with French Medical Words." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2643-2648. Portorož, Slovenia : European Language Resources Association (ELRA).

Guilbert, Louis. 1973. "La Spécificité Du Terme Scientifique et Technique." *Langue Française*, no. 17 : 5-17.

Humbley, John. 2016. "Catherine Resche (Dir.), Terminologie et Domaines Spécialisés, Approches Plurielles. Paris : Classiques Garnier, Rencontres 143, Série Linguistique 2, 2015." *ASp. La Revue Du GERAS*, no. 70 : 127-132.

Inkpen, Diana, T Sima Paribakht, Farahnaz Faez, and Ehsan Amjadian. 2016. "Term Evaluator : A Tool for Terminology Annotation and Evaluation." *International Journal of Computational Linguistics and Applications* 7 (2).

Jones, Karen Spärck. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval." *Journal of Documentation*.

Kageura, Kyo, and Bin Umino. 1996. "Methods of Automatic Term Recognition : A Review." *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 3 (2) : 259-289.

Kanoulas, Evangelos, Rene Spijker, Dan Li, and Leif Azzopardi. 2018. "CLEF 2018 Technology Assisted Reviews in Empirical Medicine Overview." In *CLEF 2018 Evaluation Labs and Workshop : Online Working Notes. CEUR-WS, France*, 1-20.

Ley, Philip. 1988. *Communicating with Patients : Improving Communication, Satisfaction and Compliance.* Croom Helm.

Nakagawa, Hiroshi. 2001. "Experimental Evaluation of Ranking and Selection Methods in Term Extraction." *Bourigault D, L'Homme M.-C., Jacquemin C.(Éd.), Recent Advances in Computational Terminology, John Benjamins Publishing Company*, 303-26.

Planas, Emmanuel. 2012. "BiTermEx Un Prototype d'extraction de Mots Composés à Partir de Documents Comparables via La Méthode Compositionnelle (BiTermEx, A Prototype for the Extraction of Multiword Terms from Comparable Documents through the Compositional Approach) [in French]." In Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, Volume 2 : TALN, 415-422. Grenoble, France : ATALA/AFCP.

Robertson, Stephen, Hugo Zaragoza, and others. 2009. "The Probabilistic Relevance Framework : BM25 and Beyond." *Foundations and Trends® in Information Retrieval* 3 (4) : 333-389.

Roche Mathieu. 2018. "Définition pluridisciplinaire de la notion de "terme"". In : TOTh 2017. Terminologie et ontologie : Théories et applications. Roche Christophe (ed.). Chambéry : Université Savoie Mont Blanc, 63-72.

Salton, Gerard, and Chung-Shu Yang. 1973. “On the Specification of Term Values in Automatic Indexing.” *Journal of Documentation* 29 (4) : 351-372.

Sandoval, Antonio Moreno, Julia Díaz, Leonardo Campillos Llanos, and Teófilo Redondo. 2019. “Biomedical Term Extraction : NLP Techniques in Computational Medicine.” *IJIMAI* 5 (4) : 51-59.

Simon, Nisha Ingrid, and Vlado Kešelj. 2018. “Automatic Term Extraction in Technical Domain Using Part-of-Speech and Common-Word Features.” In *Proceedings of the ACM Symposium on Document Engineering 2018*, 51. ACM.

Tran, Thi Mai, H Chekroud, P Thiery, and A Julienne. 2009. “Internet et Soins : Un Tiers Invisible Dans La Relation Médecine/Patient.” *Ethica Clinica* 53 : 34-43.

Vecchiato, Sara, and Sonia Gerolimich. 2013. “La Langue Médicale Est-Elle « trop Complexe » ?” *Nouvelles Perspectives En Sciences Sociales : Revue Internationale de Systémique Complexe et d’études Relationnelles* 9 (1) : 81-122.

Ventura, Juan Antonio Lossio, Clement Jonquet, Mathieu Roche, and Maguelonne Teisseire. 2014. “Towards a Mixed Approach to Extract Biomedical Terms from Text Corpus.” *International Journal of Knowledge Discovery in Bioinformatics (IJKDB)* 4 (1) : 1-15.

Vu, Thuy, Aiti Aw, and Min Zhang. 2008. “Term Extraction through Unithood and Termhood Unification.” In *Proceedings of the Third International Joint Conference on Natural Language Processing : Volume-II*.

Abstract

In this study, we propose a different perspective on the concept of weight of technical terms by focusing on the notion of “technicality” as a semantic property of the linguistic unit itself. The basic idea is that the value of technicality of a term is inversely proportional to its polysemic nature. We formalise the v-tech formula and carry out an experimental evaluation in order to 1) compare the v-tech value with other collection-based measures of termhood, and 2) integrate the v-tech formula in the score of a retrieval model for systematic reviews task for the medical domain.