Sélection Robuste de Mesures de Similarité Sémantique à partir de Données Incertaines d'Expertise

Stefan Janaqi, Sébastien Harispe, Jacky Montmain, Sylvie Ranwez

Sélection Robuste de Mesures de Similarité Sémantique à partir de Données Incertaines d'Expertise

Stefan Janaqi[1], Sébastien Harispe[1], Jacky Montmain[1], Sylvie Ranwez[1],

[1] Centre de Recherche LGI2P/Ecole des mines d'Alès
Parc scientifique G. Besse, 30035 Nîmes cedex 1, France.
prenom.nom@mines-ales.fr

Résumé :
L'exploitation d'ontologies pour la recherche d'information, la découverte de connaissances ou le raisonnement approché nécessite l'utilisation de mesures sémantiques qui permettent d'estimer le degré de similarité entre des entités lexicales ou conceptuelles. Récemment un cadre théorique abstrait a été proposé afin d'unifier la grande diversité de ces mesures, au travers de fonctions paramétriques générales. Cet article propose une utilisation de ce cadre unificateur pour choisir une mesure. A partir du (i) cadre unificateur exprimant les mesures basées sur un ensemble limité de primitives, (ii) logiciel implémentant ce cadre et (iii) benchmark d'un domaine spécifique, nous utilisons une technique d'apprentissage semi-supervisé afin de fournir la meilleure mesure sémantique pour une application donnée. Ensuite, sachant que les données fournies par les experts sont entachées d'incertitude, nous étendons notre approche pour choisir la plus robuste parmi les meilleures mesures, *i.e.* la moins perturbée par les erreurs d'évaluation experte. Nous illustrons notre approche par une application dans le domaine biomédical.

Mots-clés:
Cadre unificateur, robustesse de mesures, incertitude d'expert, mesures de similarité sémantique, ontologies.

Abstract:
Knowledge-based semantic measures are cornerstone to exploit ontologies not only for exact inferences or retrieval processes, but also for data analyses and inexact searches. Abstract theoretical frameworks have recently been proposed in order to study the large diversity of measures available; they demonstrate that groups of measures are particular instantiations of general parameterized functions. In this paper, we study how such frameworks can be used to support the selection/design of measures. Based on (i) a theoretical framework unifying the measures, (ii) a software solution implementing this framework and (iii) a domain-specific benchmark, we define a semi-supervised learning technique to distinguish best measures for a concrete application. Next, considering uncertainty in both experts' judgments and measures' selection process, we extend this proposal for robust selection of semantic measures that best resists to these uncertainties. We illustrate our approach through a real use case in the biomedical domain.

Keywords:
unifying framework, measure robustness, uncertain expertise, semantic similarity measures, ontologies.

1 Introduction

Les ontologies, définies comme des conceptualisations formelles, partagées et explicites [2], ont montré leur efficacité dans différents domaines où elles servent de support à la recherche d'information, l'inférence de connaissances ou la recherche inexacte. Elles sont au cœur de nombreuses applications académiques ou industrielles, qui associent l'expertise spécifique d'un domaine à un système à base de connaissance (analyse des gènes, systèmes de recommandation, systèmes de recherche d'information…). Elles jouent un rôle important dans la découverte de connaissances, qui repose sur des techniques de raisonnement approximatif et pour cela utilise des *mesures sémantiques*.

Ces mesures sont des fonctions estimant le degré de similarité de concepts définis dans une ontologie [9, 13]. Par extension, elles permettent d'estimer la proximité sémantique de ressources (*e.g.* maladies) indexées par des concepts (*e.g.* syndromes) [6]. Parmi la grande diversité de mesures sémantiques (voir [4]), les mesures de similarité sémantique à base de connaissances (SSM) sont utilisées pour comparer les significations de concepts par rapport à des indications. Les SSM sont utilisées dans un large spectre d'applications : désambiguïsation de texte, d'algorithmes d'extraction d'information, repositionnement de médicament, analyse de produits de gênes [4]. Ainsi un grand nombre de SSM ont été conçues de façon *ad hoc*, dans un domaine spécifique, et peu d'études se sont intéressées à leur comparaison. C'est pourquoi il

est très difficile de sélectionner une mesure adaptée à un usage spécifique.

Dans la continuité de contributions consacrées à la généralisation de mesures [1, 11, 13], nous avons récemment proposé un cadre théorique unificateur des SSM [3]. Ce cadre fournit une décomposition des SSM selon un petit nombre de primitives et nous avons montré que la plupart des SSM sont des instances de familles de fonctions paramétrées exploitant ces primitives.

Ce résultat théorique ouvre également des perspectives concernant les applications pratiques des mesures : définition de nouvelles mesures, interprétation, analyse de sensibilité, sélection de la meilleure mesure dans un contexte donné… Dans ce but, le cadre unificateur théorique, a été complété par le développement d'une librairie logicielle, la SML (*Semantic Measures Library*) [5] (http://www.semantic-measures-library.org). Ce programme générique, disponible en *open source* permet de mettre en œuvre et tester un grand nombre de SSM disponibles.

Le cadre unificateur théorique et la SML peuvent ainsi être utilisés afin de répondre aux questions fondamentales se rapportant aux mesures sémantiques : quelle mesure utiliser pour une application concrète ? Est-ce que il y a des groupes de mesures ayant le même comportement ? Quelles conséquences derrière le choix d'une mesure précise ?

Dans ce papier, nous nous intéressons à l'incertitude qui peut affecter la sélection d'une mesure spécifique dans un contexte particulier. Nous considérons qu'un ensemble de couples de concepts (x, y) est donné et que les similarités attendues $sim(x, y)$ sont fournies par des experts du domaine. Nous proposons d'utiliser ces données, le cadre unificateur théorique ainsi que la librairie SML afin de sélectionner la 'meilleure' mesure. Des techniques d'apprentissage semi-supervisé sont utilisées pour identifier les primitives et les paramètres d'une famille de mesures qui permettent de respecter au mieux les évaluations de similarité fournies par les experts du domaine. Ensuite, nous étudions l'impact de l'incertitude dans les évaluations expertes de la base d'apprentissage sur le résultat de l'identification. Quelle est la validité de la mesure identifiée s'il est admis que les experts puissent se tromper en exprimant une valeur de similarité précise ? Nous définissons un modèle simple de l'incertitude experte dans les évaluations de la base d'apprentissage. Nous nous intéressons ensuite à l'impact des erreurs d'estimation sur l'identification d'une mesure précise. Nous montrons que pour des niveaux d'incertitude 'raisonnables', il existe un large choix de mesures, qui sans être des solutions optimales au problème d'identification initial (*i.e.*, supposé sans erreur d'estimation), donnent des résultats comparables.

Ce papier présente un premier travail sur une chaine complète de traitement de l'information nécessaire à l'identification robuste de la meilleure mesure associée à un jeu de données expertes. Chacune des étapes de ce traitement peut être réexaminée en fonction de l'expression des incertitudes disponible dans les jeux de données. Les benchmarks dont nous disposons ne donnent pas d'indication sur l'incertitude des jugements d'experts. En pratique, les évaluations discrètes fournies sont le résultat de la fusion d'un collectif d'experts… Disposer des évaluations de chacun des experts aurait pu nous conduire à la construction de distribution de possibilités pour l'évaluation collective comme nous l'avons proposé dans [16] et aurait conduit à une chaine de traitement possibiliste pour la gestion d'incertitudes dans l'identification de la mesure.

Le papier est organisé comme suit: la section 2 présente le cadre unificateur des SSM ; la section 3 fournit une procédure d'apprentissage pour choisir les primitives et les paramètres à partir des données d'experts. Un modèle simple d'incertitude sur les estimations d'experts est proposé. Un indicateur numérique de robustesse est fourni (ainsi que sa visualisation) ; la section 4 illustre l'application de notre approche dans le domaine biomédical ; la section 5 apporte des conclusions et ouvre de nouvelles perspectives à ces travaux.

2 Définition de SSM au travers d'un cadre unificateur.

Cette section introduit les mesures de similarité sémantiques pour la comparaison de couples de concepts définis dans une ontologie – le lecteur intéressé par plus de détails peut se référer à [4]. Nous présentons ensuite les grandes lignes du cadre abstrait unificateur (introduit dans [3]), qui permet d'exprimer à la fois les mesures existantes et d'en définir de nouvelles. Puisque nous nous limitons à la similarité sémantique, nous considérons la réduction G d'une ontologie réduite à $\preccurlyeq$, un préordre sur la taxonomie des concepts C de l'ontologie. Les notations suivantes sont utilisées :

- $A(u)$: ensemble des ancêtres du concept u, *i.e.*, $A(u) = \{c \mid u \preccurlyeq c\}$.
- $depth(u)$: longueur du plus court chemin de u à la racine de la taxonomie.
- $LCA(u, v)$ – Least Common Ancestors : ensemble des ancêtres communs à u et v, les plus spécifiques.
- $IC(u)$ – contenu informationnel de u : défini à l'origine par $-\log(p(u))$, où $p(u)$ est la probabilité d'utilisation du concept u. Des alternatives topologiques ont été proposées [4].
- $MICA(u, v)$ – Most Informative Common Ancestor : ancêtre commun à u et v qui a le plus grand contenu informationnel.

Il est montré dans [3], que la définition d'un large ensemble de SSM peut se ramener au choix d'un petit ensemble de primitives. Nous introduisons ici les expressions juste nécessaires à la définition de l'apprentissage semi-supervisé présenté dans la section suivante. Le cadre unificateur est composé de deux éléments principaux : un ensemble de primitives et des familles paramétrées de fonctions définissant comment les primitives peuvent être associées pour construire des mesures. Nous notons par $\mathbb{K}$ un domaine contenant tout sous-ensemble de la taxonomie (*e.g.*, $A(u)$, G_u^+ sous-graphe de G induit par $A(u)$). Les primitives utilisées par le cadre unificateur sont :

- **Représentation Sémantique** (**ρ**): Une forme canonique adoptée pour représenter un concept $\rho : C \to \mathbb{K}$. On note $\rho(u)$ ou $\tilde{u}$, *la représentation sémantique* du concept u. Par exemple, un concept peut être vu comme l'ensemble des significations qu'il couvre, *i.e.*, l'ensemble $D(u)$ de ses descendants dans G.
- **Spécificité d'un concept** (**θ**): La spécificité $\theta(u)$ d'un concept u est estimée par une fonction $\theta : C \to \mathbb{R}^+$, telle que $u \preccurlyeq v \Rightarrow \theta(u) \geq \theta(v)$. Le contenu informationnel *IC* d'un concept est un exemple de fonction θ.
- **Spécificité de la représentation d'un concept** (**Θ**). Le degré de spécificité de la représentation d'un concept $\tilde{u}$, $\Theta(\tilde{u})$, est estimé par une fonction $\Theta : \mathbb{K} \to \mathbb{R}^+$. Θ vérifiant, $u \preccurlyeq v \Rightarrow \Theta(\tilde{u}) \geq \Theta(\tilde{v})$ (si $\forall u, \rho(u) = id(u)$, alors $\mathbb{K} = C$ et $\Theta = \theta$).
- **Commonalité** (**Ψ**); La *commonalité* (points communs) de deux concepts (u, v), par rapport à leurs représentations $(\tilde{u}, \tilde{v})$ est donnée par une fonction $\Psi(\tilde{u}, \tilde{v})$, $\Psi : \mathbb{K} \times \mathbb{K} \to \mathbb{R}^+$.
- **Différence** (**Φ**): La quantité de connaissance présente en $\tilde{u}$ et non présente en $\tilde{v}$ est estimée par une fonction $\Phi(\tilde{u}, \tilde{v})$: $\Phi : \mathbb{K} \times \mathbb{K} \to \mathbb{R}^+$.

Les fonctions abstraites ρ, Ψ, Φ sont des primitives que l'on peut introduire pour réécrire et interpréter une majorité de SSM. Ici, nous donnons sim_{RM}, une formulation abstraite du '*ratio model*' introduit par Tversky [15]

en utilisant différentes acceptions de ces primitives, présentées dans la Table 1 ci-dessous.

$$sim_{RM}(u, v) = \frac{\Psi(\tilde{u}, \tilde{v})}{\alpha\, \Phi(\tilde{u}, \tilde{v}) + \beta\, \Phi(\tilde{v}, \tilde{u}) + \Psi(\tilde{u}, \tilde{v})} \qquad \textbf{(1)}$$

Cas	1	2	3	4
$\rho(u) = \tilde{u}$	G_u^+	$A(u)$	$A(u)$	$A(u)$
$\Theta(\tilde{u})$	$depth(u)$	$IC(u)$	$\sum_{c \in A(u)} IC(c)$	$\lvert A(u) \rvert$
$\Psi(\tilde{u}, \tilde{v})$	$depth(LCA(u,v))$	$IC(MICA(u,v))$	$\sum_{c \in A(u) \cap A(v)} IC(c)$	$\lvert A(u) \cap A(v) \rvert$

Table 1. Exemples d'instances des primitives définies par le cadre unificateur.

Mesures	Cas	Paramètres
$sim_{Wu\ \&\ Palmer}(u,v) = \frac{2\ depth(LCA(u,v))}{depth(u)+depth(v)}$	1	$\alpha = 0.5,\ \beta = 0.5$
$sim_{Lin}(u,v) = \frac{2\ IC(MICA(u,v))}{IC(u)+IC(v)}$	2	$\alpha = 0.5,\ \beta = 0.5$
$sim_{Faith}(u,v) = \frac{IC(MICA(u,v))}{IC(u)+IC(v)-IC(MICA(u,v))}$	2	$\alpha = 1, \quad \beta = 1$
$sim_{Mazandu}(u,v) = \frac{2\ \sum_{c \in A(u) \cap A(v)} IC(c)}{\sum_{c \in A(u)} IC(c)+\sum_{c \in A(v)} IC(c)}$	3	$\alpha = 0.5, \quad \beta = 0.5$
$sim_{CMatch}(u,v) = \frac{\lvert A(u) \cap A(v) \rvert}{\lvert A(u) \cup A(v) \rvert}$	4	$\alpha = 1, \quad \beta = 1$

Table 2. Réécriture de quelques SSM définies pour le calcul de similarités entre.

3 Sélection des primitives et des paramètres à utiliser pour la définition d'une mesure

L'objectif d'un concepteur de SSM est de définir la 'bonne' combinaison des paramètres $(\rho, \Theta, \Psi, \Phi, \alpha, \beta)$ pour définir la mesure abstraite la mieux adaptée à son contexte si l'on s'en tient au *ratio model*, sim_{RM}. Les étapes de la sélection sont :

Etape 1. Choisir une liste finie de primitives $(\rho, \Theta, \Psi, \Phi), \Pi = \{\pi_l \mid \pi_l = (\rho_l, \Theta_l, \Psi_l, \Phi_l), l = 1, \dots, L\}$ selon des considérations sémantiques et l'application en vue (voir table 1).

Etape 2. Choisir les paramètres $(\alpha_l, \beta_l), l = 1, \dots, L$ à appliquer avec π_l dans le ratio model. Soit une liste finie de SSM est disponible (voir table 2) et il suffit alors de calculer l'écart sur la base d'apprentissage entre les similarités fournies par les experts et les similarités calculées avec chacune des mesures candidates ; la mesure qui produit l'erreur cumulée minimale est conservée ; soit la recherche du couple (α_l, β_l) est vue comme un problème d'optimisation paramétrique continu. C'est ce point de vue qui est développé ici.

Pour un choix $(\pi_l, \alpha_l, \beta_l)$, nous notons, pour tout couple de concepts $(x,y) : s_l(x,y) \equiv sim_{RM(\pi_l, \alpha_l, \beta_l)}(x,y)$. Supposons maintenant que les experts aient fourni les similarités pour un

ensemble de N couples de concepts (x_k, y_k) : $s_k = s(x_k, y_k),\ k = 1, \dots, N$. Notons $\mathbf{s} = [s_1, \dots, s_N]^T$ le vecteur de ces valeurs de similarité expertes. On peut estimer la qualité d'une SSM s_l par sa capacité à reproduire les estimations d'experts. Notons $\mathbf{s}_l$ le vecteur de coordonnées $\mathbf{s}_l(k) = s_l(x_k, y_k), k = 1, \dots, N$. Les valeurs $s_l(x_k, y_k)$ dépendent de (α_l, β_l). Il est donc possible de trouver un couple (α_l^0, β_l^0) qui permette de s'approcher au mieux des estimations expertes, par exemple le couple qui optimisera la corrélation entre $\mathbf{s}$ and $\mathbf{s}_l$:

$$\begin{cases} \max_{\alpha_l, \beta_l} corr(\mathbf{s}, \mathbf{s}_l(\alpha_l, \beta_l)) \\ 0 \leq \alpha_l, \beta_l \leq M \end{cases} \qquad (P_l)$$

Il est facile de montrer qu'il existe une constante M, telle que la solution optimale du problème (P_l) soit dans le domaine $0 \leq \alpha_l, \beta_l \leq M$.

Par ailleurs, les estimations des experts sont entachées d'incertitude. Dans le domaine des applications biomédicales, les réponses des experts sont souvent des valeurs sur une échelle linéaire discrète : $v_i = v_0 + i.\Delta,\ i = 1, \dots, V, \Delta$ une constante (dans l'exemple montré dans la section suivante $s_k \in \{1, 2, 3, 4\}$). Les réponses des experts peuvent s'écrire sous la forme : $t_k = s_k + \varepsilon_k$, où s_k représente la vraie valeur et ε_k une erreur d'estimation. Soit ε le vecteur erreur de coordonnées ε_k. Nous supposons que $\varepsilon_k \in \{-\Delta, 0, \Delta\}$ avec une distribution de probabilité : $p(\varepsilon_k = 0) = q,\ p(\varepsilon_k = -\Delta) = p(\varepsilon_k = \Delta) = \frac{1-q}{2}$. Autrement dit, l'erreur d'estimation par un expert ne saurait dépasser une unité Δ. Les erreurs d'estimation d'une évaluation à l'autre sont supposées indépendantes (il est facile de trouver une similarité entre l'éléphant d'Asie et l'éléphant d'Afrique, elle est plus difficile à voir entre le daman et l'éléphant). Le niveau global de l'incertitude peut être contrôlé par la valeur de q. A un vecteur ε est associé un couple $(\alpha_l(\varepsilon), \beta_l(\varepsilon))$ solution du problème d'optimisation (P_l) avec $\mathbf{s} + \varepsilon$ comme vecteur des estimations expertes.

Tout choix de q définit une distribution de probabilité sur les couples $(\alpha_l(\varepsilon), \beta_l(\varepsilon)) \sim D^q_{\alpha_l, \beta_l}$ solutions de (P_l) induite par la distribution d'entrée sur l'erreur ε. La forme analytique de $D^q_{\alpha_l, \beta_l}$ n'est pas connue.

Rappelons que nous cherchons une SSM robuste : le couple optimal $(\alpha_l^0, \beta_l^0) = (\alpha(0), \beta(0))$ solution de (P_l) fournira donc une mesure pertinente seulement si celle-ci reste satisfaisante lorsque l'incertitude de l'expert obéit à la loi de distribution de paramètre q. Pour y répondre, nous générons aléatoirement des vecteurs ε, nous résolvons le problème (P_l) avec $\mathbf{s} + \varepsilon$ pour trouver $(\alpha_l(\varepsilon), \beta_l(\varepsilon))$. La distribution $D^q_{\alpha_l, \beta_l}$ est ainsi estimée empiriquement pour un « grand » nombre de vecteurs ε.

On définit ensuite l'ensemble de niveau :

$L_r = \left\{ (\alpha_l(\varepsilon), \beta_l(\varepsilon)), corr\left(\mathbf{s}, \mathbf{s}_l(\alpha_l(\varepsilon), \beta_l(\varepsilon))\right) \geq r \right\}$, *i.e.* l'ensemble des couples $(\alpha_l(\varepsilon), \beta_l(\varepsilon))$ pour lesquels la mesure $\mathbf{s}_l$ rend compte des estimations expertes initiales ($\varepsilon = 0$) de façon satisfaisante.

$R(q) = \iint_{L_r} D^q_{\alpha_l, \beta_l} d\alpha_l d\beta_l$ estime alors la proportion des couples $(\alpha_l(\varepsilon), \beta_l(\varepsilon))$ qui appartiennent à L_r. La robustesse du modèle de similarité défini par le couple (α_l^0, β_l^0) est donc d'autant plus grande que R est grand (pour un q donné). La valeur de l'intégrale double ci-dessus, peut par exemple être calculée par une méthode de type Monte Carlo.

L'existence de $D^{q}_{\alpha,\beta}$ offre une autre possibilité pour choisir un couple de paramètres robustes $(\alpha^{*}_{l}, \beta^{*}_{l})$ pour un modèle d'incertitude de paramètre q : la médiane des couples $(\alpha(\varepsilon), \beta(\varepsilon))$ pondérés par $D^{q}_{\alpha,\beta}$. Une approximation de ce point est fournie par la médiane des points générés par la méthode Monte Carlo.

Dans la section suivante, l'approche est appliquée à un cas réel : $(\alpha^{*}_{l}, \beta^{*}_{l})$ coïncide avec $(\alpha^{0}_{l}, \beta^{0}_{l})$ pour $q = 0$ (cas évident), $(\alpha^{*}_{l}, \beta^{*}_{l}) \in L_r$ pour des valeurs modérées de $1 - q$ et peut différer sensiblement de $(\alpha^{0}_{l}, \beta^{0}_{l})$ pour de grandes valeurs de $1 - q$.

4 Choix d'une SSM robuste

Les tests présentés ici s'appuient sur la librairie logicielle SML [5]. Nous avons utilisé le jeu de test proposé par Pedersen *et al*. [9], souvent cité en référence dans le domaine biomédical pour évaluer les mesures sémantiques par rapport au jugement humain. Naturellement, tous les algorithmes et traitements qui reposent sur ces mesures (*e.g.* extraction d'information, analyse de données) visent une forte corrélation avec les estimations humaines de la similarité [7–9]. Ainsi, les mesures sont évaluées de par leur capacité à imiter l'appréciation de la similarité par des experts. Le jeu de test de Pedersen *et al*. contient 29 couples de termes liés au domaine biomédical; pour chaque couple de termes, le couple de concepts correspondant est extrait de l'ontologie biomédicale SNOMED-CT [14]. Pour chaque couple, la similarité est obtenue en moyennant les évaluations données par les experts (la distribution des évaluations n'est pas disponible). Les évaluations s_k sont exprimées dans $\{1, 2, 3, 4\}$ et $\varepsilon_k \in \{-1, 0, 1\}$.

L'approche est appliquée avec chacune des instances π_l de la table 1 avec le ratio model $sim_{RM}(u, v)$ de l'équation (Eq.1). Les valeurs optimales $(\alpha^{0}_{l}, \beta^{0}_{l})$ solutions du problème (P_l) et les valeurs de corrélation correspondantes sont fournies dans la table 3. Les meilleurs résultats correspondent aux mesures des cas 2 et 4 de la table 1. On conserve seulement ces deux mesures pour la suite.

Les ensembles de niveau L_r sont représentés sur la figure 1 pour les cas 2 et 4. La probabilité qu'un expert donne une réponse erronée est $1 - q \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. Pour chaque valeur de q, 10 000 vecteurs ε sont générés aléatoirement (sur la figure, $1 - q$ est fixé à 0.1 et 0.4). Les valeurs estimées de la robustesse $R(q)$ et les couples $(\alpha^{*}_{l}, \beta^{*}_{l})$ sont fournis dans la table 4. Le modèle $(\alpha^{0}_{l}, \beta^{0}_{l})$ du Cas 2 présente non seulement une meilleure corrélation (Table 3) mais il est aussi le plus robuste (Table 4). Si on choisit $r = 0.73$ comme seuil de corrélation pour définir l'ensemble des couples L_r qui fournissent un modèle de similarité satisfaisant, il est utile pour les praticiens/concepteurs de SSMs de visualiser L_r et d'évaluer la robustesse de leur modèle afin qu'il puisse estimer la qualité de leur modèle, mais aussi sa sensibilité aux erreurs d'estimation et par suite au choix de $(\alpha^{*}_{l}, \beta^{*}_{l})$.

5 Conclusions et perspectives

Face à une très grande diversité de mesures sémantiques, il est important pour les utilisateurs de SSM d'avoir des outils pour choisir les mesures les plus adaptées à un domaine et une application spécifique. Notre approche ouvre la perspective de création de tels outils en permettant de comparer et d'évaluer les SSM dans d'autres contextes d'application.

$$sim_{RM-C2}(u, v) = \frac{IC(z)}{18.62\,(IC(u) - IC(z)) + 4.23(IC(v) - IC(z)) + IC(z)} \quad (2)$$

Table 3. Performance optimale des mesures.

	Cas 1 (C1)	Cas 2 (C2)	Cas 3 (C3)	Cas 4 (C4)
Max. Correlation Optimale (α_l^0, β_l^0)	0.719 (9.89,1.36)	**0.768** (18.62,4.23)	0.736 (7.26,0.40)	0.759 (6.17,0.77)

Table 4. Variation de la solution optimale en fonction de l'incertitude.

	$1-q$ $= 0.1$	$1-q$ $= 0.2$	$1-q$ $= 0.3$	$1-q$ $= 0.4$	$1-q$ $= 0.5$
$R_{C2}(u)$ $(\alpha_{C2}^*, \beta_{C2}^*)$	0.83 (18.62, 4.23)	0.70 (18.62, 4.23)	0.56 (15.31, 4.23)	0.49 (16.70, 4.07)	0.39 (13.71, 4.02)
$R_{C4}(u)$ $(\alpha_{C4}^*, \beta_{C4}^*)$	0.76 (6.17,0.77)	0.54 (6.17,0.76)	0.46 (5.52,0.71)	0.39 (5.12,0.64)	0.35 (4.06,0.70)

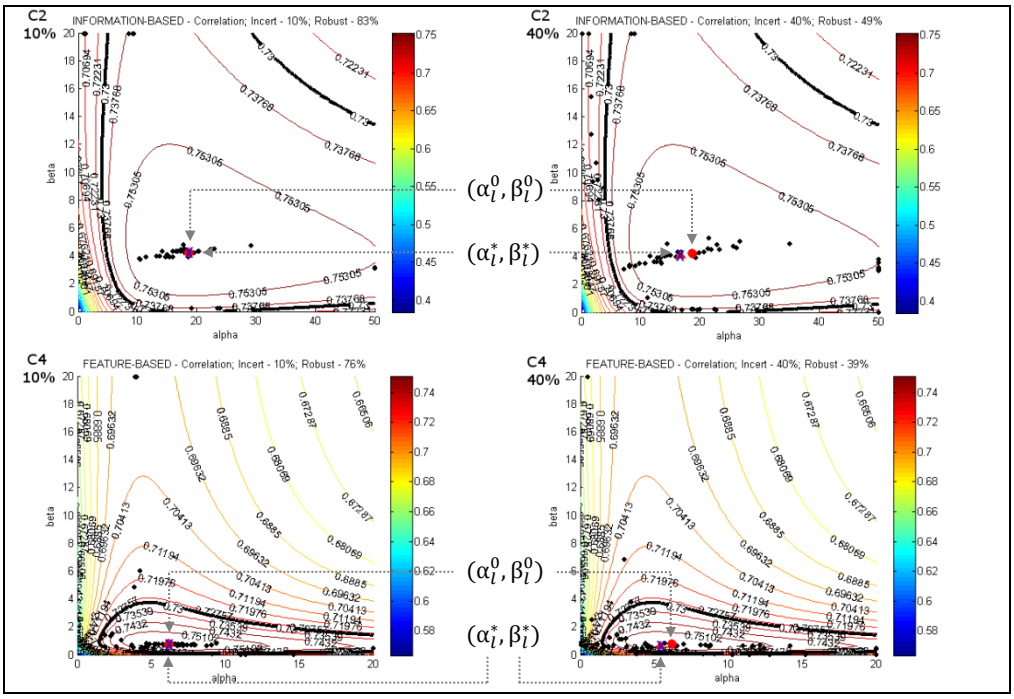

Figure 1. Robustesse des cas 2 et 4 pour une incertitude à 10% et 40%. Pour chaque figure, les couples (α_l^0, β_l^0), (α_l^*, β_l^*) et L_r sont pointés.

Ce papier se concentre sur la sélection de SSM robustes dans un contexte de données incertaines fournies par des experts humains du domaine étudié. A notre connaissance, nous sommes les premiers à proposer cette approche.

Nous proposons un estimateur analytique de la robustesse, et sa contrepartie graphique afin de caractériser cette propriété importante des SSM. Ainsi, en mettant en évidence les limites des estimateurs actuels des mesures, spécialement lorsque l'incertitude affecte les données, notre proposition ouvre des perspectives intéressantes pour la caractérisation et la sélection adéquate des mesures pour des études de domaines spécifiques. Cette approche très calculatoire s'appuie sur la librairie logicielle SML (http://www.semantic-measures-library.org).

Néanmoins, les observations concernant l'application sont basées sur l'analyse d'une configuration spécifique de mesures, utilisent une seule ontologie et un seul jeu de test. Sans remettre en cause l'approche, il sera nécessaire

d'élargir cette étude à d'autres jeux de test afin d'obtenir des conclusions plus générales. Aussi, la sensibilité de notre approche aux différentes propriétés des benchmarks (ex. la taille, la représentativité, …) reste à étudier. Nous souhaitons ainsi contribuer à une meilleure compréhension et analyse des SSM, en particulier identifier le rôle et les connexions entre l'expression abstraite des mesures, les primitives, les instances et les paramètres (ex. α, β). Enfin, notons qu'une approche similaire peut servir pour étudier des SSM suivant des expressions abstraites autres que sim_{RM}, ouvrant ainsi l'étude à une large diversité de mesures non présentées dans ce papier.

Références

1. Blanchard, E. et al.: A generic framework for comparing semantic similarities on a subsumption hierarchy. 18th Eur. Conf. Artif. Intell. 20–24 (2008).
2. Gruber, T.: A translation approach to portable ontology specifications. Knowl. Acquis. 5.2, April, 199–220 (1993).
3. Harispe, S. et al.: A Framework for Unifying Ontology-based Semantic Similarity Measures: a Study in the Biomedical Domain. J. Biomed. Inform., (2013).
4. Harispe, S. et al.: Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis. ArXiv. (2013).
5. Harispe, S. et al.: The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. Bioinformatics. in press, (2013).
6. Mathur, S., Dinakarpandian, D.: Finding disease similarity based on implicit semantic similarity. J. Biomed. Inform. 45, 2, 363–371 (2012).
7. Pakhomov, S. et al.: Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. AMIA Annu. Symp. Proc. 2010, 572–576 (2010).
8. Pakhomov, S.V.S. et al.: Towards a framework for developing semantic relatedness reference standards. J. Biomed. Inform. 44, 2, 251–65 (2011).
9. Pedersen, T. et al.: Measures of semantic similarity and relatedness in the biomedical domain. J. Biomed. Inform. 40, 3, 288–99 (2007).
10. Pesquita, C. et al.: Semantic similarity in biomedical ontologies. PLoS Comput. Biol. 5, 7, 12 (2009).
11. Pirró, G., Euzenat, J.: A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. Proceedings of the 9th International Semantic Web Conference ISWC 2010. pp. 615–630 Springer (2010).
12. Sánchez, D. et al.: Ontology-based information content computation. Knowledge-Based Syst. 24, 2, 297–303 (2011).
13. Sánchez, D., Batet, M.: Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. J. Biomed. Inform. 44, 5, 749–759 (2011).
14. Spackman, K.A.: SNOMED CT milestones: endorsements are added to already-impressive standards credentials. Healthc. informatics Bus. Mag. Inf. Commun. Syst. 21, 9, 54– 56 (2004).
15. Tversky, A.: Features of similarity. Psychol. Rev. 84, 4, 327–352 (1977).
16. Imoussaten A., Montmain J., Mauris, G., A multicriteria decision support system using a possibility representation for managing inconsistent assessments of experts involved in emergency situations. International Journal of Intelligent Systems, **29** (1), pp. 50-83, 2014.