Apprentissage de mesures de similarité sémantiques : étude d'une variante de la mesure InfoSimba

Guillaume Cleuziou and Gaël Dias

Abstract Dans cette étude nous nous intéressons aux mesures de similarité sémantiques utilisées pour évaluer la proximité (sémantique) entre des mots du langage naturel. Plus précisément nous revenons sur l'hypothèse "fausse" qui consiste à considérer comme similaires des mots souvent utilisés dans les mêmes contextes (fenêtres contextuelles, paragraphes, documents entiers). La mesure InfoSimba [3] révise en effet cette hypothèse en considérant que deux mots sont proches si leurs vecteurs de contextes (composés des mots physiquement proches) sont similaires. Nous envisageons ici d'explorer une technique de construction itérative de la mesure InfoSimba en réévaluant à chaque itération le vecteur de contexte qui décrit chaque mot.

1 Introduction

Les mesures de similarité sémantiques construites à partir de l'analyse des cooccurrences des mots dans les textes se fondent sur l'hypothèse (de Harris) suivante : deux mots sont d'autant plus proches sémantiquement qu'ils sont souvent utilisés dans les mêmes contextes [4]. Ainsi deux mots seront considérés comme très similaires s'ils apparaîssent souvent "l'un à côté de l'autre" (fenêtre contextuelle) ou dans les mêmes documents et de façon complémentaire s'ils apparaîssent rarement l'un sans l'autre. Les mesures du type : Probabilité conditionnelle (SCP), coefficients de Dice et de Jaccard ou encore l'Information Mutuelle utilisent ce principe [2], nous les qualifierons par la suite de *mesures d'association*.

Guillaume Cleuziou
LIFO-Université d'Orléans - France e-mail: guillaume.cleuziou@univ-orleans.fr

Gaël Dias HULTIG-Université de Beira Interior, Covilhã - Portugal
e-mail: ddg@di.ubi.pt

En observant que deux synonymes sont rarement utilisés conjointement, les mesures d'association précédentes ne permettent pas de rapprocher deux mots en relation de synonymie, alors que par définition même de cette relation sémantique ils devraient incarner l'exemple parfait de mots très similaires. Pour intégrer cette problématique de la synonymie, Dias [3] ou encore Cleuziou [1] utilisent la notion de "vecteur de contexte" : un tel vecteur caractérise chaque mot par l'ensemble des k mots les plus associés au sens des mesures précédentes. Dias propose alors la mesure InfoSimba qui consiste à comparer deux à deux chacun des éléments de deux vecteurs de contexte pour en déduire la similarité sémantique entre deux mots. Cleuziou étudie simplement le coéfficient de corrélation entre deux lignes de la matrice de "similarité" obtenue par une mesure d'association. Finalement l'hypothèse sous-jacente (simplifiée) est que deux mots sont similaires sémantiquement si les mots auxquels ils sont souvent associés sont similaires.

Dans la suite de l'article nous rappelons brièvement la mesure InfoSimba puis proposons une variante récursive de cette mesure avant d'en présenter une première évaluation quantitative.

2 La mesure InfoSimba

La mesure InfoSimba s'appuie sur une mesure d'association (par exemple la mesure SCP) d'une part pour établir un vecteur de contexte pour chaque mot w_i et d'autre part pour comparer deux vecteurs de contextes. Étant donné un vocabulaire V, un vecteur de contexte $\mathbf{w_i}$ d'un mot $w_i \in V$ est construit en recherchant les k mots les plus associés à w_i selon la mesure SCP (avec k un paramètre fixé).

$$\mathbf{w_i} = (w_{i,1}, w_{i,2}, \ldots, w_{i,k}) \text{ où } w_{i,j} = \underset{w \in V \setminus \{w_{i,1},\ldots,w_{i,j-1}\}}{\operatorname{argmax}} SCP(w_i, w)$$

Une fois établis, les vecteurs de contextes sont en fait utilisés comme des ensembles (l'ordre des mots dans le vecteur importe peu). Un opérateur de comparaison entre ensemble (la mesure InfoSimba) est défini ainsi :

$$InfoSimba(\mathbf{w_i}, \mathbf{w_j}) = \frac{\sum_{p=1}^{k} \sum_{q=1}^{k} X(w_{i,p}) X(w_{j,q}) SCP(w_{i,p}, w_{j,q})}{norm}$$

Dans cette définition $X(.)$ est une fonction de pondération des mots (combinaison de $tf \times idf$ et d'une notion de densité[1]) et *norm* est un coefficient normalisateur permettant à la mesure InfoSimba de prendre ses valeurs dans $[0,1]$.

L'hypothse que la mesure InfoSimba tente de formaliser est la suivante :

(H1) *Deux mots sont sémantiquement proches si les mots auxquels ils sont souvent associés sont eux-mêmes sémantiquement proches*.

On peut noter le caractère récursif de cette proposition. Dans un soucis de simplification la mesure InfoSimba formalise plutôt l'hypothèse suivante :

[1] voir l'article de Dias et Alves [3] pour plus de précisions.

(H2) *Deux mots sont sémantiquement proches si les mots auxquels ils sont souvent associés sont eux-mêmes souvent associés.*

Afin de mieux correspondre à l'intuition de base de la mesure InfoSimba nous en proposons une version récursive que nous définissons de la manière suivante :

$$SimbaRec_N(\mathbf{w_i}, \mathbf{w_j}) = \frac{\sum_{p=1}^{k} \sum_{q=1}^{k} SimbaRec_{N-1}(\mathbf{w_{i,p}}, \mathbf{w_{j,q}})}{norm}$$

Ainsi à chaque itération, la similarité entre deux mots est calculée à partir des similarités calculées à l'itération précédente. Nous proposons d'initialiser le processus ($SimbaRec_0$) par la mesure InfoSimba originelle présentée précédemment.

3 Évaluation de la mesure

Pour évaluer l'intérêt de la mesure, nous avons constitué un vocabulaire composé de 38 mots-clés extraits d'articles scientifiques issus des trois sources de publications suivantes :

Sources	Mots-clés
Conférence LREC (2000)	Annotation Guidelines, Bracketed Corpus, Chinese Language Processing, Quality Control, Combining Systems, Machine Learning, Tagging, Knowledge-Rich NLP, Multilingual Corpora, Parallel Corpora, POS Tagging
Conférence WWW (2002)	Content Distribution Networks, Data Consistency, Data Dissemination, Dynamic Data, HTTP, Leases, Protocol Design, Pull, Push, Scalability, TCP Splice, Web Proxy, World Wide Web
Revue JSAI (1997)	Classification Rule, Macro Rule, Concept Learning, Constructive Induction, Colored Digraph, Logic Programming, Problem Solving, Program Transformation, Unfolding, Control of Computation, Natural Language Processing, Ill-Formedness, Robust Parsing, Integration

Table 1 38 mots-clés et sources de référence.

En supposant que deux mots-clés issus d'une même source (paires dites correctes) sont d'avantage similaires sémantiquement que deux mots issus de sources différentes, nous allons observer la répartition des valeurs de similairtés entre les couples valides et non-valides. Les similarités sont calculées en observant l'utilisation des mots-clés du vocabulaire sur le Web[2].

La figure 1 présente la répartition des paires de mots correctes sur 10 quantiles, correspondant à une discrétisation de l'ensemble des valeurs de similarités en 10 sous-ensembles de mêmes tailles et triés des plus faibles valeurs (1er quantile) aux plus fortes valeurs (10ème quantile). Cette expérience a été réalisée en utilisant la mesure d'information mutuelle comme mesure d'association[3], des vecteurs de contextes de taille 10 ($k = 10$) et une pondération uniforme des mots (fonction $X(.)$).

[2] voir Cleuziou [1] pour plus de précisions.

[3] Mesure d'association sur laquelle nous avons obtenus les meilleurs résultats.

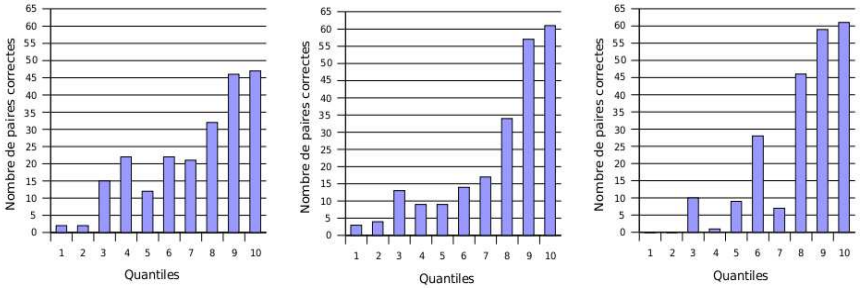

Fig. 1 Répartition des valeurs de similarité : information mutuelle (à gauche), InfoSimba (au centre), SimbaRec_1 (à droite).

Sur cette figure on observe un report progressif des paires correctes vers les plus fortes valeurs de similarité :

1. en passant de la mesure d'information mutuelle à la mesure InfoSimba utilisant cette même mesure d'association,
2. en passant de la mesure InfoSimba originelle à sa version récursive (première itération).

D'autres résultats expérimentaux sur ce vocabulaire (non présentés ici) ont montrés d'une part que seules les mesures d'Information Mutuelle et SCP permettent d'obtenir un gain de précision en utilisant la mesure SimbaRec plutôt que la mesure InfoSimba de base ; et d'autre part que l'amélioration sensible obtenue sur ces mesures se limite à la première itération.

En conclusion, malgré les limites évoquées précédemment, les résultats obtenus sur cette première expérience laissent à penser que les hypothèses formulées ne sont pas infondées. Ainsi la mesure InfoSimba peut, sous certaines conditions (taille des vecteurs de contextes, choix de la mesure d'association) mieux quantifier la similarité sémantique entre mots que les mesures d'association usuelles. De plus la variante récursive que nous avons proposée peut (sous les mêmes conditions) améliorer encore la qualité de cette mesure.

References

1. Cleuziou, G.: Une méthode de classification non-supervisée pour l'apprentissage de règles et la recherche d'information. Thèse de doctorat. LIFO, Université d'Orléans (2004)
2. Cleuziou, G., Clavier, V., Martin, L.: Une méthode de regroupement de mots fondée sur la recherche de cliques dans un graphe de cooccurrences. In: L. ENSAIS (ed.) 5èmes rencontres Terminologie et Intelligence Artificielle, pp. 179–182. Strasbourg, France (2003). Poster
3. Dias, G., Alves, E.: Unsupervised topic segmentation based on word co-occurrence and multiword units for text summarization. In: Proceedings of the ELECTRA Workshop associated to 28th ACM SIGIR Conference, pp. 41–48. Salvador, Brazil (2005)
4. Harris, Z., Gottfried, M., Ryckman, T., Mattick, P., Daladier, A., Harris, T.N., Harris, S.: The form of Information in Science: Analysis of an immunology sublanguage. Dordrecht : Kluwer Adademic Publishers (1989)