PPMI : étude formelle d'une variante à valeurs positives de la PMI

François Role*, Mohamed Nadif**

*Université Paris Descartes. Département informatique. 143, avenue de Versailles, 75016
francois.role@univ-paris5.fr,
**LIPADE, Université Paris Descartes, 45, rue des Saints-Pères, 75005
mohamed.nadif@parisdescartes.fr

1 Introduction

Dans de nombreuses tâches allant de l'expansion de requêtes à l'extraction de terminologie ou la construction d'ontologies, il est crucial de pouvoir déterminer statistiquement le degré d'association sémantique entre deux mots x et y dans un corpus. Parmi les nombreuses mesures d'association disponibles (test du rapport de vraisemblance, test du Khi-deux, etc.), l'information mutuelle spécifique *Pointwise Mutual Information* notée pmi a été très utilisée en lexicographie à partir du début des années 1990 (travaux de Church et Hanks), notamment pour l'extraction de paires de mots ayant tendance à apparaître fréquemment ensemble (collocations). Depuis cette époque, on a cependant souvent reproché à la pmi d'une part de favoriser les mots ayant une basse fréquence et d'autre part de ne pas prendre de valeurs dans un intervalle borné.

Définie par $\log \frac{p(x,y)}{p(x)p(y)}$, cette mesure a effectivement tendance à attribuer des scores d'association très élevés à des paires impliquant des mots rares puisque le dénominateur est petit dans ce cas. Par ailleurs, contrairement à $MI(X,Y)$ qui est la moyenne des pmi et qui est toujours positive, la $pmi(x,y)$ peut être positive ou négative et a une valeur nulle en cas d'indépendance complète entre deux mots x et y.

Des variantes empiriques ont été proposées pour remédier à ces deux problèmes. Parmi les variantes les plus utilisées, on peut citer celles dites de la "famille PMI^k" (Daille, 1994). Elles consistent à élever empiriquement au carré ou au cube le numérateur apparaissant dans la définition de la pmi, ce qui donne, en prenant l'exemple du carré, $pmi^2(x,y) = \log \frac{p(x,y)^2}{p(x)p(y)}$. Cette correction produit un rééquilibrage en faveur des mots fréquents mais dans des proportions non indiquées dans la littérature. Par ailleurs pour atténuer le second défaut de la pmi (valeurs variant entre $-\infty$ et $-\log p(x,y)$), certains auteurs préconisent de ne pas tenir compte des valeurs négatives. Nous proposons donc de définir sur des bases plus formelles les corrections à apporter.

2 Une interprétation formelle des variantes de la pmi

Une dérivation, qui ne peut être reproduite ici, nous a permis de montrer que la correction apportée par la pmi^2 consiste en fait à retirer à la pmi de deux mots x, y la quantité $-\log p(x, y)$ qui représente l'information (au sens de la théorie de l'information) associée à la paire x, y. On sait que cette information est d'autant plus faible que la paire est plus fréquente.

La correction par soustraction de la quantité $-\log p(x, y)$. ne permet cependant pas d'obtenir des valeurs normalisées. Pour ce faire, certains auteurs ont donc proposé de la diviser par $-\log p(x, y)$ (Gerlof, 2009), valeur qui se trouve également être la borne supérieure de la pmi. Cette mesure, dite NPMI (*Normalized PMI*) aboutit également à relever le score des paires à fréquence élevée et pourrait être qualifiée de "rééquilibrage par division". Des expérimentations poussées et menées sur de larges volumes de texte extraits de la Wikipedia nous ont cependant montré que la correction apportée est relativement faible et que des paires à fréquence élevée peuvent être repoussées assez loin. L'autre limite importante de la NPMI est que contrairement à ce que son nom pourrait laisser penser, elle prend ses valeurs dans $[-1, 1]$. Rappelons au passage que ce problème de non normalisation est partagé par les mesures pmi^k qui prennent, elles, leurs valeurs entre $-\infty$ et 0. Une variante permettant de régler, simultanément et en connaissance de cause, les deux problèmes semble donc être la mesure que nous appelons $ppmi$ (*positive pmi*) :

$$ppmi = 2^{pmi(x,y)-(-\log p(x,y))}$$

On obtient ainsi une variante positive de la pmi dont le type de correction peut être interprété dans le contexte de la théorie de l'information. Par ailleurs, avec cette mesure, il est possible de quantifier précisément l'impact des fréquences sur les scores obtenus. En effet si la pmi de deux paires diffère de $k >= 0$, c'est-à-dire si $pmi(z, t) = pmi(x, y) + k$, on peut montrer que :

$$\begin{cases} ppmi(z,t) = ppmi(x,y) & \text{si} \quad \log \frac{p(x,y)}{p(z,t)} = k \\ ppmi(z,t) > ppmi(x,y) & \text{si} \quad \log \frac{p(x,y)}{p(z,t)} < k \\ ppmi(z,t) < ppmi(x,y) & \text{si} \quad \log \frac{p(x,y)}{p(z,t)} > k \end{cases} \tag{1}$$

Références

Daille, B. (1994). Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques. *PhD thesis, Université Paris 7 (1994).*

Gerlof, B. (2009). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, Gunter Narr Verlag, pp. 31–40. Chiarcos, Eckart de Castilho & Stede (eds).

Summary

In this paper, we present a normalized variant of the pmi (Pointwise Mutual Information) whose behavior and corrrection factors are formally defined.