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.