Longueur de branches et arbres de mots

Philippe Gambette[1], Nuria Gala[2], Alexis Nasr[2], Alain Guénoche[3]

Résumé

Les arbres permettent de visualiser les mots les plus fréquents d'un texte en reflétant leur co-occurrence dans ce texte. Pour cela, les mots sont placés aux feuilles de l'arbre, et leur fréquence d'apparition commune (dans une même phrase, un même paragraphe...) est indiquée par la distance entre les deux mots dans l'arbre, c'est-à-dire la somme des longueurs des arêtes de l'arbre sur le chemin qui va de l'un à l'autre. La construction de ces arbres par des méthodes de classification arborée, parfois inspirées de la phylogénie, conduit généralement à des arbres dont les arêtes qui mènent aux feuilles ont une longueur beaucoup plus grande que les arêtes internes. Pour éviter ce phénomène qui nuit à la lisibilité de l'arbre, il est usuel de choisir des longueurs identiques pour toutes les arêtes de l'arbre, au détriment de la fidélité entre la distance représentée par l'arbre et la distance entre mots qui a servi à le construire. Nous proposons une approche qui assure la lisibilité de l'arbre tout en maintenant ses qualités informatives. Pour cela, nous comparons plusieurs méthodes pour attribuer, après la construction de l'arbre, des longueurs à ses branches, de telle manière que ces longueurs soient compatibles avec une interprétation de l'arbre comme un ensemble de classes de mots. Nous les évaluons sur des données de distances sémantiques entre mots d'une même famille morphologique, calculées notamment à partir de co-occurrences dans le TLFI. Nous illustrons également la meilleure méthode retenue sur un corpus textuel.

Mots clés : textométrie, visualisation, arbre de mots, nuage arboré, co-occurrence, réseau sémantique, partition.

1. Introduction

Les arbres de mots se sont ajoutés aux projections et aux réseaux de co-occurrence parmi les outils développés pour l'analyse textométrique des textes (Mayaffre, 2008). Ils permettent en effet de représenter de manière esthétique un nombre limité de classes de mots emboîtées (linéaire en le nombre de mots), tout en faisant varier les tailles de caractères des mots, par exemple dans les nuages arborés (Gambette & Véronis, 2009).

Toutefois, contrairement aux arbres phylogénétiques dont les longueurs de branches représentent une distance évolutive entre espèces (Felsenstein, 2004), il est difficile d'interpréter les longueurs de branche d'un arbre de mots calculées comme une approximation de la distance entre les mots. A défaut de respecter ces longueurs, c'est la topologie de l'arbre qui est privilégiée. Celui-ci est donc interprété de la manière suivante : chaque arête interne correspond à une bipartition entre deux groupes de mots. Plutôt qu'effectuer une lecture globale de l'arbre, on y recherche donc des classes de mots regroupés dans un même sous-arbre, séparé du reste de l'arbre par une longue arête. Ce type de lecture est troublé par un inconvénient des méthodes de construction d'arbres inspirées de la bioinformatique : la longueur de branches menant aux feuilles (appelées *arêtes externes*). En effet, en raison de la structure très particulière des formules de co-occurrence de mots (Evert, 2005), on peut constater expérimentalement que la longueur des arêtes internes d'un nuage arboré est souvent très petite par rapport à celle des arêtes menant aux feuilles, ce qui réduit la lisibilité de ses sous-arbres.

Pour éviter ce phénomène, nous proposons de recalculer les longueurs de branches après la construction de l'arbre, de telle manière qu'elles assurent sa lisibilité, tout en facilitant la lecture de l'arbre comme une partition en classes de l'ensemble des mots. Pour cela, nous proposons d'utiliser des formules proposées par Guénoche & Garreta (2002) pour évaluer la qualité des arêtes d'un arbre. Ces formules indiquent si les deux ensembles de mots séparés par une arête sont effectivement bien séparés d'après la matrice de distance. Ainsi, on attribuera à chaque arête une longueur proportionnelle à son score de qualité.

Afin d'évaluer la pertinence de ces formules, nous proposons un algorithme qui construit une partition d'un ensemble de mots, à partir d'un arbre de mots, en découpant successivement ses arêtes dans l'ordre décroissant des longueurs, comme illustré en Figure 1. Nous testons cet algorithme sur des partitions de familles de mots de la base de données Polymots (Gala & Rey, 2008), en comparant les partitions obtenues par cet algorithme sur des distances de mots basées sur leur co-occurrence dans le TLFI avec des partitions sémantiques réalisées manuellement (Gala et al., 2011).

[1] Université Paris-Est – LIGM, 5 boulevard Descartes, 77454 Champs-sur-Marne, philippe.gambette@univ-mlv.fr

[2] Université Aix-Marseille – LIF, 163 av. de Luminy, case 901, 13288 Marseille Cedex 9, {nuria.gala, alexis.nasr}@lif.univ-mrs.fr

[3] CNRS – IML, Campus de Luminy, 163 av. de Luminy, case 907, 13288 Marseille Cedex 9, guenoche@iml.univ-mrs.fr

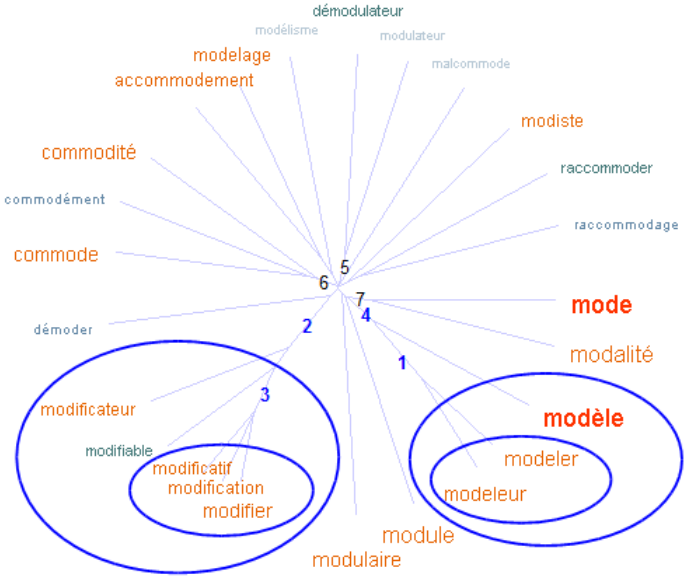

Figure 1: Nuage arboré des mots de la famille de « mode » dans la base de données Polymots (Gala & Rey, 2008), construit par les logiciels TreeCloud (Gambette & Véronis, 2009) et SplitsTree (Huson & Bryant, 2006). Les 7 arêtes les plus longues sont numérotées dans l'ordre des longueurs décroissantes, et les classes de mots correspondant aux arêtes 1 à 4 sont entourées.

2. La classification arborée comme source d'une partition sémantique

Le concept de base de l'interprétation d'un arbre comme partition d'un ensemble de mots consiste à choisir un ensemble de sous-arbres disjoints, et à interpréter leurs ensembles de feuilles comme une partition de l'ensemble des mots de l'arbre. Si cette interprétation se fonde sur les longueurs de branches, alors les sous-arbres sont créés en découpant chaque arête interne, dans l'ordre des longueurs décroissantes, jusqu'à un critère d'arrêt. Ainsi, en considérant l'ensemble des ensembles de mots de chacun de ces sous-arbres, on obtient une partition de l'ensemble des mots.

Pour ré-évaluer la longueur d'une arête de l'arbre, plusieurs formules sont possibles à partir d'une matrice de distances entre mots de l'arbre (Guénoche & Garreta, 2002) : taux de liens, taux de bons triplets, taux de bons quadruplets, taux d'accord des paires, ratio des longueurs moyennes... Par exemple, le taux de bons triplets d'une arête consiste à déterminer, pour tout ensemble de trois mots $\{a,b,c\}$, où a et b sont situées d'un côté de l'arête et c de l'autre côté, la proportion de triplets qui vérifient $d(a,b) \leq \min(d(a,c),d(b,c))$, d étant la distance sémantique entre les mots.

Pour chacune de ces formules de longueur des arêtes, nous obtenons une partition de l'ensemble des mots placés aux feuilles en considérant les composantes connexes induites par l'effacement des k arêtes les plus longues de l'arbre.

3. Réseau et classes sémantiques comme corpus d'évaluation

La base de données Polymots comprend 20 000 mots regroupés en 2 000 familles de mots centrées autour d'un mot racine. Vingt de ces familles ont été partitionnées manuellement en classes sémantiques. Par exemple, voici la partition pour un extrait de la base correspondant à la famille du radical « mode » : [accommodement, commode, commodément, commodité, malcommode] [démoder, mode, modiste] [modalité] [modelage, modeler, modeleur] [modèle, modélisme] [modifiable, modificateur, modification, modificatif, modifier] [modulaire, module, modulateur, démodulateur] [raccommodage, raccommoder].

Pour chacune des formules de longueur d'arêtes, nous effectuons, pour chaque famille de mots, une comparaison, à l'aide de l'indice de Rand corrigé, entre les partitions construites manuellement et celles construites à partir d'une distance sémantique prenant en compte les co-occurrences des mots dans le TLFi ainsi que leur nombre d'affixes communs, en utilisant l'algorithme de découpage des arêtes par longueur décroissante. Nous obtenons de cette façon un score de qualité pour chacune des formules de calcul des longueurs d'arêtes de l'arbre. L'ensemble de cette méthodologie est montrée en Figure 2.

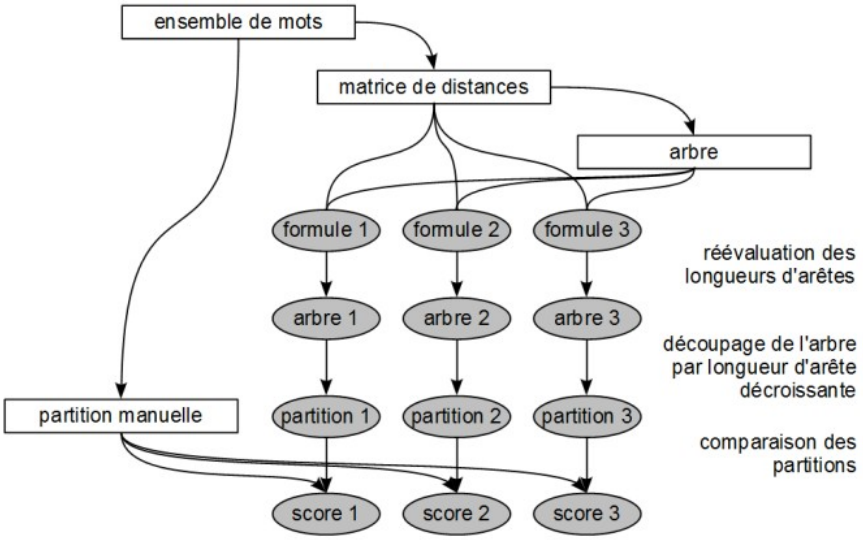

Figure 2: Méthodologie d'évaluation des formules de calcul de longueur des arêtes de l'arbre.

Références

Evert S. (2005). *The Statistics of Word Cooccurrences, Word Pairs and Collocations*. Thèse de l'Université de Stuttgart, pp. 75-91.

Felsenstein J. (2004). *Inferring Phylogenies.* Sinauer Associates.

Gala N., Hathout N., Nasr A., Rey V. & Seppälä S. (2011). Création de clusters sémantiques dans des familles morphologiques à partir du TLFi, In *Actes de TALN'11*.

Gala N. & Rey V. (2008). Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. In *Actes de TALN'08*.

Gambette P. & Véronis J. (2009). Visualising a Text with a Tree Cloud. In Locarek-Junge H. and Weihs C., éditeurs, *Classification as a Tool of Research, Proc. of IFCS'09 (11th Conference of the International Federation of Classification Societies)*.

Guénoche A. & Garreta H. (2002). Representation and Evaluation of Partitions. In *Classification, clustering and data analysis, Proc. of IFCS'02*.

Huson D.H. & Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology and Evolution*, 23(2): 254-267, logiciel disponible sur www.splitstree.org.

Mayaffre D. (2008). Quand “travail”, “famille”, “patrie” co-occurrent dans le discours de Nicolas Sarkozy. Étude de cas et réflexion théorique sur la co-occurrence. In Heiden S., Pincemin B., éditeurs, *Actes des JADT'08*.