Variante de max-cut pour extraire des entités

Introduction

Mon but est d'extraire des entités importantes de la loi. Dans la modélisation qui va suivre, ce sont uniquement des groupes nominaux, mais on peut généraliser.

Le problème des entités dans la loi, c'est qu'elles sont hiérarchiques: comme des poupées russes, elles sont composées de sous-entités.

C'est cela qui rend ma tâche difficile: est-ce que je dois extraire "santé", "agence nationale des produits de santé", "produits de santé" ... Qu'en est-il de "santé publique" ... ?

Structure du problème

J'ai mis au point une mesure (basée sur l'entropie) qui exprime à quel point on perd de l'information en retirant une sous-entité B d'une certaine entité A.

Par exemple, passer de "santé publique" à "….. publique" pourrait faire perdre 3 bits d'information.

En représentant toutes mes dépendances entre entités avec ces scores, j'obtiens un DAG pondéré:

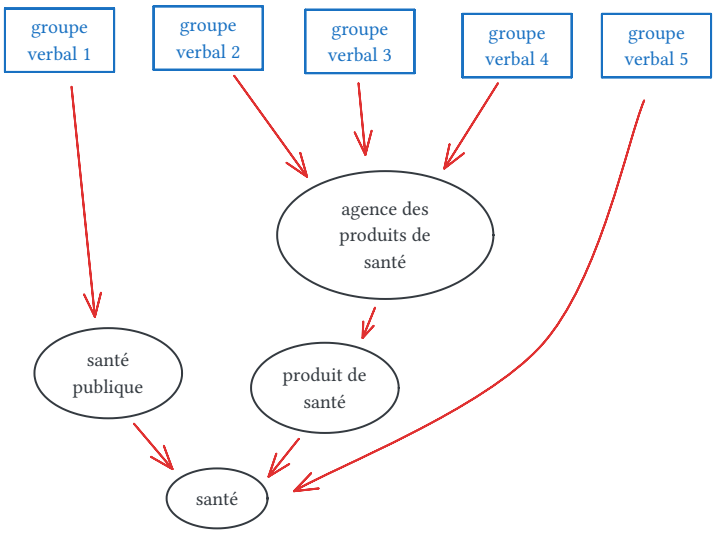

Figure 1: graphe des entités candidates

Mon but est alors d'identifier un ensemble de k noeuds du graphe, qui contient le plus d'information.

Prenons un exemple:

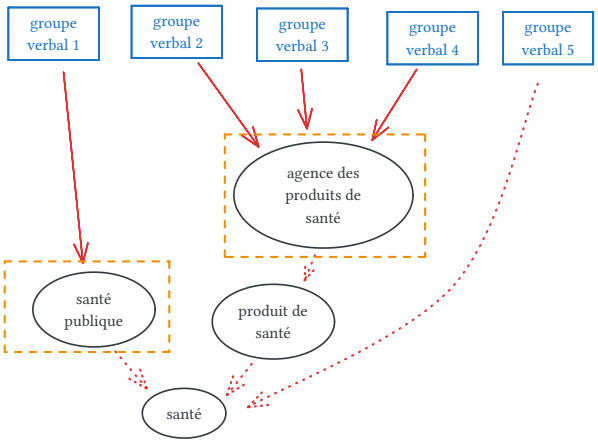

Figure 2: 2 entités indépendantes

Dans ce cas, pour compter l'information extraite, je compte le poids de toutes les arêtes qui arrivent aux candidats selectionnés.

Mais les entités ne sont pas indépendantes: si j'extrait une entité du texte, je ne peux plus extraire ses sous-entités. Exemple:

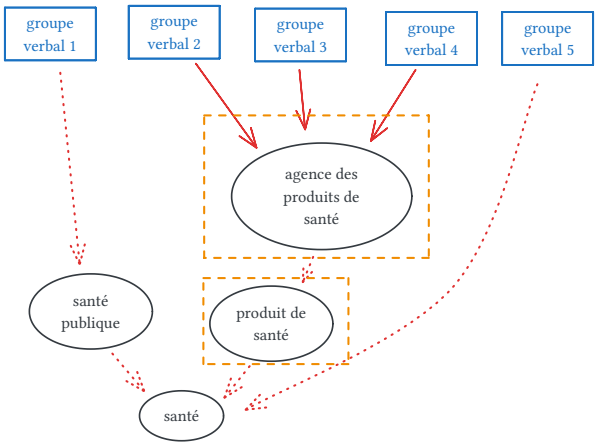

Figure 3: 2 entités dont une est sous-entité de l'autre. Il ne faut pas compter l'arête entre 'agence des produits de santé' et 'produits de santé'

Ceci correspond à une coupe: on compte toutes les arêtes telles que le départ n'est pas dans l'ensemble, et l'arrivée est dans l'ensemble.

Sous forme matricielle, avec M ma matrice des poids, je cherche à optimiser:

$$\underset{(x_i)}{\text{argmax}} \sum_{i,j} M_{ij} x_i (1 - x_j)$$

Mais ce n'est pas exactement ce que je veux. Regardons l'exemple suivant:

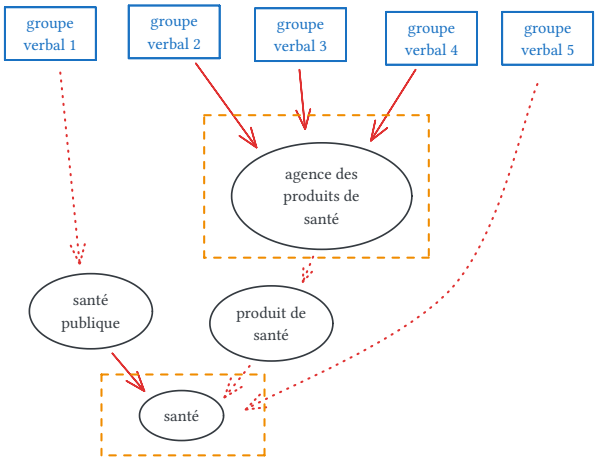

Figure 4: 2 entités non adjacentes. Il ne faut pas comptabiliser 'produit se santé' -> 'santé', mais il faut comptabiliser 'santé publique' -> santé

On veut donc comptabiliser toutes les arêtes dont qui partent d'un noeud dont aucun parent n'est selectionné, et qui arrivent à un neoud selectionné.

Sous forme maricielle:

$$\underset{(x_i)}{\text{argmax}} \sum_{i,j} M_{ij} x_i \prod_{j \text{ parent de } i} (1 - x_j)$$

En ajoutant la contraine d'un nombre maximal de sommets selectionnés:

$$\underset{(\sum x_i \leq k)}{\text{argmax}} \sum_{i,j} M_{ij} x_i \prod_{j \text{ parent de } i} (1 - x_j)$$