UID mémoire Aurelien

Aurelien Said Housseini

February 2023

Contents

1 résumé de l'expérience de Jaeger

Dans cette expérience, Jaeger mesure l'effet prédicteur d'UID sur l'omission de *that* dans les propositions complétives en anglais. Il utilise un certain nombre de prédicteurs nécessaires à la prédiction, ces prédicteurs peuvent relever de plusieurs aspects langagiers, ils ont été classés dans les catégories suivantes !1(table) !2(lister catégories de façon cohérente). Hormis ces prédicteurs il mesure UID de façon très localisée, en évaluant la surprise que causerait l'apparition d'une complétive comme objet du verbe matrice de ladite complétive.

Pour ce faire, sur la totalité du corpus, il mesure la probabilité d'apparition d'une complétive en objet d'un verbe, $P_v(completive)$ où v est le verbe matrice.

Afin de rendre mieux compte de la densité d'information, Jaeger passe par la mesure de surprise, le $-log()$ de la probabilité d'un évenement, ce qui donne comme mesure finale :

$$-log(P_v(completive))$$

!3(Comment le corpus a été obtenu/les complétives trouvées)

Jaeger a utilisé le corpus Switchboard (ref), une collection de conversations téléphoniques en anglais. Le corpus comprend beaucoup de données sur les locuteurs ainsi que des annotations en partie du discours. Il a utilisé les parties du discours pour générer un patron morpho-syntaxiques permettant de pré-sélectionner les phrases comprenant des propositions complétives, et les échantillons ont été finalement sélectionnés à la main. Cela a donné un corpus de 6716 phrases comprenant des complétives avec ou sans that.

1.1 Replication

1.1.1 Différences

Différences entre l'expérience de Jaeger et ma réplication

Le corpus switchboard annoté n'étant pas disponible. J'ai dû chercher des données assez conséquentes pour pouvoir obtenir un réel effet prédicteur. Je n'ai pas été en mesure de trouver un corpus d'anglais oral annoté ne serait-ce qu'en parties du discours. J'ai donc décidé d'utiliser des corpus d'écrit. J'ai utilisé tous les corpus d'anglais écrit disponibles en UD !4(expliquer UD) !5(décrire brièvement les corpus). Du fait de la modalité de mes corpus je n'ai pas pu retrouver les mêmes prédicteurs que Jaeger, tous ceux liés à l'oral ou aux données des locuteurs ne sont donc pas présents dans mon modèle. Afin de retrouver les autres prédicteurs je suis passé par des règles symboliques basées sur les relations syntaxiques présentes en UD (et en SUD) !5(expliquer l'apport de SUD) !6(montrer quelques/toutes les règles symboliques).

1.1.2 Modèle statistique

Explication du modele mixte Facteur

Une fois les prédicteurs extraits du corpus, l'objectif est de mesurer leur effet sur la prédiction de l'omission de that. Pour ce faire, je passe par la fonction glmer() de R, qui permet de lancer un modèle mixte !7(expliquer en détails le principe du modèle). Le passage par ce type de modèle nécessite quelques modifications de la façon d'encoder les prédicteurs : Pour les prédicteurs catégoriels il faut les rendre binaires Pour les prédicteurs pouvant etre obtenu à partir d'un autre il faut faire une regression et garder les résidus

Une fois les facteurs bien établis, il suffit de lancer le modèle pour obtenir le résultat suivant.

Les effets sont ceux qui étaient attendus dans l'article de Jaeger, !8(description des direction et de la force de prédiction) Malgré le passage à une modalité écrite il est intéressant de noter qu'UID semble tout de meme avoir un effet. !9(mieux mettre en forme).

Maintenant je possède une valeur étalon à laquelle il sera intéressant de comparer des méthodes plus actuelle de mesurer UID

2 Methodologie

2.1 corpus

Afin de réaliser la tâche de prédiction de l'omission de that dans les complétives, il est nécessaire d'avoir un corpus duquel il est facile d'extraire les complétives avec that ainsi que les complétives pour lesquelles il serait possible que that soit utilisé comme conjonction de subordination. Pour ce faire sans simplement extraire les complétives manuellement, il faut à minima un corpus annoté en catégories morphosyntaxiques, c'est pourquoi nous travaillons sur des corpus UD, annoté à la fois en catégories morphosyntaxiques et en syntaxe de dépendance. Le corpus que nous utilisons principalement est le corpus GUM, regroupant (quantité) d'énoncés/tokens d'anglais de modalités variées (liste de genre) annotés et vérifiés manuellement.

2.2 Ecrit vs Oral

Dans le cadre de ce mémoire, la majeure partie des énoncés sont issus de l'écrit. UID est un phénomène principalement étudié dans le cadre de l'oral pour ce qui est de la mesure d'UID dans la production langagière (Il existe des études mesurant l'effet d'UID dans l'écrit pour mesurer l'acceptabilité à la lecture). Cela nous permet donc d'évaluer l'effet d'UID dans de l'écrit.

2.3 Extraction des autres variables

2.4 Vectorisation des complétives

Les modèle linéaires ne prennent en entrée pour leurs prédiction que des données numériques au format assez particulier. Pour que le modèle puisse prédire

l'omission de that dans les complétives, nous procédons en réalité à une certaine forme de vectorisation des complétives, chaque variable est une dimension d'un vecteur qui correspond à une complétive, et c'est à partir de ces dimensions que le modèles pourra inférer des effets de chaque variable sur ce qu'il doit prédire : l'omission de that. (exemple du dataframe des complétives)

2.5 Selection des complétives

Certains verbes matrices ne peuvent que très rarement être gouverneur d'une complétive, ce type de verbe est assez commun et nuit assez fortement aux mesures à réaliser, en particulier celle de Jaeger, se reposant essentiellement sur la probabilité qu'un verbe puisse gouverner une complétive. Nous ne prenons donc que les complétives dont le verbe gouverneur apparait au moins une fois en tant que gouverneur dans (x) complétives et pouvant présenter des cas avec ou sans omission de that.

2.6 Illustration de UID avec une phrase

2.7 "Problèmes" avec le proxy de Jaeger (incertain de ce paragraphe, dur à mettre en mots)

Le proxy utilisé par Jaeger présente (3) problèmes, (1) il ne prend pas du tout en compte le contenu informationnel de l'énoncé, uniquement la quantité informationnelle qu'apporterait une complétive après un verbe matrice, ce qui rend ce proxy entièrement non contextuel, (2) c'est un proxy fondamentalement lexical, qui repose uniquement sur la capacité d'un verbe, (3) il est fondamentalement lié au phénomène langagier en jeu

3 Mesurer UID

3.1 utilisation de LLM

L'un des objets principaux de ce mémoire est de mesurer les capacités des LLM en tant que proxy d'UID, les LLM étant largement utilisés dans des études de modélisation cognitive computationnelle. Pour nos manipulations, nous utilisons GPT-2, le modèles renvoyant des logits très facilement transformés en probabilités.

3.2

Afin d'exploiter au mieux l'effet de l'uniformité sur la sélection d'une construction, je vais principalement travailler sur la différence d'uniformité entre l'énoncé produit (avec ou sans *that*) et un énoncé fictif présentant l'alternative.
Pour ce faire je vais donc traiter des doublons pour chaque proposition complétive.

Exemple de doublon :

My boss confirmed that we were absolutely crazy
My boss confirmed we were absolutely crazy

Au final, j'ai distingué 3 façons de considérer UID, plus ou moins localement :

1 Ecart entre la quantité d'information apportée et une quantité d'information plus globale (pouvant aller de la quantité d'information moyenne de la phrase, à la quantité moyenne d'information de la langue)

2 Minimisation de l'écart entre deux quantité d'information juxtaposées

3 L'Entropie

3.3 Généralités

On définit :

- s_i : une phrase contenant au moins une proposition complétive (un échantillon du corpus)
- D_{ij} : un doublon d'énoncés avec ou sans *that*, introduisant la j^{eme} complétive de la i^{eme} phrase
- d^+ij : la version de la phrase avec *that* introduisant la j^{eme} complétive
- d^-ij : la version de la phrase sans *that* introduisant la j^{eme} complétive
- $I(t_k)$: la quantité d'information pour le $k^{eme} aptinstalltexlive - sciencetexlive - latex - extratexlive - extra - utilslatexmktexlive - publisherstexlive - science$ token d'une séquence
- $\delta(t_k)$: Saut d'informativatité entre deux tokens voisins $I(t_k) - I(t_{k-1})$
- λ : l'indice de la position du premier token d'une complétive

Le but est de trouver la version de la phrase la plus uniforme. Pour cela, on utilise des scores d'uniformité u^+ij et u^-ij respectivement attribués à d^+ij et d^-ij. Le score de préférence de l'omission de *that* est calculé comme suit :

$$\Omega(D_{ij}) = u^-ij - u^+ij \tag{1}$$

$\Omega(D_{ij})$ est non borné. Si $\Omega(D_{ij}) > 0$ d'après l'hypothèse d'UID, il devrait y avoir une préférence pour l'omission de *that* ; autrement, une préférence pour son inclusion.

3.4 Point de vue global

Alternative:

- Utilisation de la variance/écart type de la quantité d'information
- On peut mesurer l'écart de la quantité informationnelle moyenne de chaque élément d'un doublon à la quantité informationnelle moyenne du document ou de la langue

3.4.1 Variance

La valeur attribuée à d^+ est la variance de la quantité d'information pour chaque token; idem pour d^-.

Pour ce qui est de l'utilisation de la variance (ou de l'écart-type), le score suivant suffit.

$$\frac{1}{\sigma^2} \tag{2}$$

3.4.2 Ecart avec une valeur d'uniformité

Le score attribué à d^+ (resp. d^-) est l'écart d'informativité au point λ (θ_n, où n correspond au nombre de tokens à droite de λ utilisés pour établir l'informativité à ce point) à un score d'uniformité de référence (noté $\bar{x}$).

$$\theta_n = \frac{1}{n} \sum_{i=\lambda}^{\lambda+n} I(t_i) \tag{3}$$

Le score d'uniformité d'un d s'obtient comme suit :

$$u(d_{ij}) = |\theta_n - \bar{x}| \tag{4}$$

Les différents scores d'uniformité possibles sont:

- l'informativité moyenne calculée sur l'ensemble des tokens du contexte gauche (**??**)
- l'informativité moyenne sur l'ensemble des tokens de la phrase ()
- l'informativité moyenne sur l'ensemble du document ()

$$i - gauche_k = \frac{1}{k} \sum_{i=k}^{n-1} I(t_i) \tag{5}$$

$$i - phrase = \frac{1}{|s|} \sum_{i=1}^{|s|} I(t_i) \tag{6}$$

$$i - doc = \frac{1}{|doc|} \sum_{i=1}^{|doc|} I(t_i) \tag{7}$$

3.5 Point de vue local

Le score d'uniformité local est basé sur les écarts internes à la phrase. Ici, les mesures étant pour la plupart des facteurs de désuniformité, le score en est l'inverse.

Scores à considérer:

- inverse de l'amplitude (amplitude = écart entre la plus haute et la plus basse valeur d'informativité)
- inverse de la taille du plus grand saut (saut = différence d'informativité entre deux tokens voisins)
- inverse de la moyenne des écarts entre tokens adjascents pour une phrase

L'amplitude est l'écart entre la densité d'information la plus élevée et la densité d'information la plus faible dans la phrase :

$$u(d) = \frac{1}{amplitude} \tag{8}$$

L'écart maximum $\Delta(s_i)$, est la différence de densité d'information la plus élevée entre deux token voisins :

$$u(d) = \frac{1}{\max_{2 \leq k \leq |s|} \delta(t_k)} \tag{9}$$

La taille moyenne des sauts $\bar{\delta}(s_i)$, est la moyenne des écarts entre token adjacents pour une phrase :

$$\frac{1}{\frac{1}{|s_i|} \sum_{k=1}^{|s|} \delta(t_k)} \tag{10}$$

3.6 Entropie

Pour l'entropie il faut considérer comme variable aléatoire, l'omission de that, pour cela il faut donc trouver un moyen de donner une probabilité aux deux issues possibles.

Il est possible d'imaginer comparer la somme des logits des tokens obtenus à partir du CC onset. Par exemple si *that this is so sweet* donne une somme de valeur a et *this is so sweet* une somme de valeur b alors on pourrait imaginer $P^{+}_{that} = \frac{a}{a+b}$ et $P^{-}_{that} = \frac{b}{a+b}$. On pourrait donc utiliser cette probabilité pour calculer l'entropie de Shannon.