Modèles en génération automatique de textes

Par **Jean-Pierre BALPE**
Université Paris VIII

Pour faire percevoir l'ensemble de la problématique, je vais commencer par un niveau élémentaire, celui de la génération de mots, démarche très utile notamment pour la publicité qui est sans cesse confrontée à la possibilité de création automatique de noms de marques.

Soit un ensemble de 5 lettres: combien de «mots» possibles puis-je créer avec elles. Par exemple: b, a, l, p, e.

Cela dépend.

Une infinité si je ne crée aucun modèle de contraintes puisqu'un mot peut alors être une seule lette : «a» par exemple ou une combinaison quelconque dans n'importe quel ordre de ces lettres : «abbalpepapeebaleppea…» etc. Pour créer un mot qui soit susceptible d'être accepté comme tel dans une langue naturelle, il faut se donner un minimum de règles:

- longueur maximale du mot (sachant, par exemple que la longueur moyenne en français est de 7 caractères)
- longueur minimale du mot
- règle de répétition ou non-répétition
- règle «d'orthographe», différente par exemple en français, en italien ou dans n'importe quelle autre langue utilisant un alphabet latin: «ea» est acceptable en anglais, rare en français, interdit en italien, etc…
- règle d'existence dans une langue donnée

Mais, même avec toutes ces règles on obtient un nombre important de mots. En français, depuis «a», «le», «la» jusqu'à «albe» ou «abel». Par contre n'existe aucun terme français contenant les 5 lettres choisies. Si j'ajoute une lettre «r» ou «s» ou «i» je découvre bien sûr davantage de résultats certains utilisant les 5 lettres initiales: «laper», «parle», «pales», «sable», «bipale», «pibale», etc… Mais je ne peux obtenir ces résultats que par un «apport extérieur» au modèle, un dictionnaire de langue, ici le français.

Autrement dit, il n'existe pas de modèle abstrait susceptible de décider que «pabler» n'est pas français, mais que «laper» l'est.

Pour le dire autrement et m'amuser d'une métaphore empruntée à la physique, ce niveau des éléments ultimes de constitution du français est porteur d'une forte entropie et, pour réduire cette entropie, la langue va devoir le soumettre à d'importantes contraintes.

Ce simple exemple doit suffire pour affirmer qu'une modélisation que j'appellerai de «haut niveau» c'est-à-dire ne s'appuyant que sur des modèles abstraits et quelques règles n'est que difficilement envisageable. Cela semble aujourd'hui évident. Pourtant depuis un siècle, depuis la fondation du structuralisme linguistique par Ferdinand de Saussure, puis à travers d'approches comme les linguistiques Chomskyennes, Harrissiennes ou les travaux de chercheurs comme Meehan, Schank, Danlos, etc c'est ce type d'approche qui a dominé.

Le niveau des mots est un niveau simple si l'on se contente d'ignorer leur sens car, comme l'affirme déjà Saussure avec la notion d'arbitraire du signe, disant qu'il n'existe aucun rapport naturel entre le signifié (le concept) et le signifiant (l'image acoustique), en d'autre termes entre le sens et sa réalisation visuelle et acoustique (le mot). Cet arbitraire du signe explique que, pour désigner un même concept (par exemple "soleil"), il soit possible d'utiliser différentes

réalisations graphiques et phoniques, telles que sun en anglais, sol en portugais, sonne en allemand, شمس chems en arabe, et cætera.

Qu'en est-il maintenant des niveaux supérieurs sur lesquels se sont concentrés la plupart des recherches linguistiques: ceux des phrases et des textes?

Si nous faisons la même expérience et choisissons cinq mots au hasard: «avodiré, cisaillement, involution, vêture, recharge», il est facile de s'apercevoir que leur combinaison, qu'elle soit, ne donnera jamais une phrase française. Il faut donc compliquer un peu le modèle en introduisant ce que l'on appelle la syntaxe avec des notions de types: adjectif, substantif, verbe notamment. On peut alors jouer à faire du Henri Michaux: «recharge cisaillement avodiré vêture involution», ce serait mieux avec une virgule (nouvelle notion, celle de rythme linguistique) : «recharge cisaillement avodiré, vêture involution», ce serait mieux encore en introduisant d'autres notions, celle de conjugaison par exemple: «recharger cisaillement avodiré, vêture involution». Peu à peu «de la langue» apparaît ainsi : «recharger un cisaillement avodiré, vêture l'involution» et elle se révèle davantage si l'on triche encore un peu: «demain moi travailler gâteau papou». C'est encore du galimatias mais l'apparition d'une syntaxe minimale et de relations de sens rend la chaîne acceptable.

Le nombre de combinaisons possibles de mots français est infini mais, comme le montre l'exemple ci-dessus, peu d'entre elles sont acceptables. L'entropie y est encore considérable qui ne pourra être réduite que par l'imposition de liens et de structures à lous les constituants élémentaires des langues.

La question est donc quels modèles faut-il se donner pour qu'un algorithme qui, comme l'on sait est un modèle de raisonnement formel, permette d'approcher de l'acceptable?

Pour mieux comprendre, inversons la perspective en partant de trois chaînes incontestablement recevables en français standard:

1. La cavale au valaque avala l'eau du lac et l'eau du lac lava la cavale au valaque
2. Un dernier verre est parfois le dernier
3. Une habitante de Lunéville a porté plainte contre un dermatologue après le suicide de son fils à qui elle avait prescrit un générique du Roaccutane.

Les questions posées étant:

1. Quels sont les mécanismes mentaux mis en jeu pour la compréhension de ces chaînes
2. Quelle probabilité y a-t-il d'obtenir ces deux chaînes par des modèles formels?

Chacune des ces phrases, tout en étant parfaitement compréhensible, pose des problèmes de natures différentes: phonétiques, articulatoire, encyclopédiques, expérience sociétale, connaissances externes, etc qui en rendent l'interprétation, par une informatique algorithmique, donc déductive, presque impossible.

Mais essayons d'abord à un niveau plus simple: un substantif et un adjectif, cas dans lequel cet adjectif sera réalisé par «grand»

Un homme grand
Un grand homme
Un homme, grand
Un homme, un grand
Une grande femme
Une grande dame
Un grand mouvement
La grande demoiselle
Un grand bal
Un grand moment

Un grand film
Un grand cheval
Un grand terrain

Quel que soit le niveau auquel on se place on retrouve ce même genre de difficultés qui montrent que la syntaxe — ici relativement facile à modéliser — est sans grande importance, que c'est la sémantique qui mène le bal et que cette sémantique s'appuie largement sur des conventions acceptées ou des connaissances externes à la langue. Cette difficulté est évidemment encore plus grande si l'on ne se contente pas d'une seule langue.

Il n'y a pas de langue sans locuteur. L'impossibilité de déchiffrer des langues disparues comme le Rongo Rongo de l'île de Pâques dont nous avons pourtant des hiéroglyphes, en est un exemple. C'est le couple écriture-lecture qui va tenter de réduire l'entropie au maximum au point de risquer de figer la langue dans des constructions qui se veulent définitives et où ne peut plus entrer aucun jeu. Mais le texte-cristal est une visée presque impossible à atteindre, si ce n'est dans quelques sous-langages très spécialisés, aussi, le lecteur va-t-il jouer un rôle essentiel en réduisant l'entropie par l'interprétation. Le lecteur doit ainsi faire preuve de coopérativité lectorielle, cette attitude est essentielle pour rendre possible la génération automatique.

Ces constatations conduisent à penser qu'il est nécessaire d'adopter une démarche totalisante reposant presque entièrement sur le recours à la mémoire des multiples enregistrements que propose Internet en prenant comme position de départ que presque tout « a déjà été dit ». Cette approche propose comme modèle de base la génération la juxtaposition de données existantes.

Cependant une démarche à la Google qui consisterait à engranger des milliards de données, devrait pour les utiliser à bon escient et, surtout, pour les réunir en des textes acceptables, n'est encore pas satisfaisante car elle demande des modélisations sémantiques, une démarche à la Pierre Lévy par exemple. En effet, un ensemble d'éléments textuels, même bien formés syntaxiquement ne formera que très rarement un texte, car leur combinatoire introduit un autre niveau d'entropie. Si je tire cinq phrases au hasard dans trois ouvrages différents :

1. Une table était dressée sous un arbre devant la maison.
2. Pour trouver la réponse, Ptolémée avait demandé à ses serviteurs de mesurer soigneusement les dimensions extérieures du bâtiment.
3. N'hésitez pas à commenter votre code.
4. Je t'assure que ce chapeau est amoureux de tes cheveux.
5. Une ville dans un petit état de l'Ouest.

Même en essayant de les combiner de toutes les façons possibles, on n'obtient pas un texte cohérent: un texte est en effet basé sur des relations sémantiques complexes entre éléments constitutifs.

Le problème est donc, pour la génération automatique de voir comment peuvent être modélisées certaines des contraintes sémantiques indispensables et à quel niveau elles doivent être introduites. Par exemple, des données comme :

1. un cheval mange de l'avoine
2. l'Aloe Vera est une des plantes médicinales des plus puissantes et des plus connues.

Peuvent à l'évidence être décrites dans deux registres différents: 1.description animale, 2.description botanique qui, à un premier niveau suffiront à ce qu'un générateur de texte ne les traite de façon indistincte.

Suivant les besoins, ces descriptions peuvent alors être plus ou moins sophistiquées en introduisant des classes ou des sous-classes. Celles-ci pouvant appartenir à un domaine donné ou se voulant plus généralistes. La recherche d'une description conceptuelle générale étant par exemple, le but que se donnent les travaux de Pierre Lévy.

Cependant, une fois la classification sémantique des données établie, celle-ci s'avère vite insuffisante car un texte n'est pas la simple juxtaposition de phrases appartenant au même domaine. Un texte repose aussi sur des relations internes dont il faut tenir compte. Il est alors tout à fait possible de faire générer des textes en choisissant les bons niveaux de modélisation. Le niveau inférieur à celui de la donnée brute est celui que j'appellerai « graphes de connaissances» qui tente également de diminuer l'entropie naturelle des données linguistiques. Il repose en partie sur la notion de «graphe de connaissance» (GC).
Un «graphe de connaissance» est uns structure établissant des liens entre éléments de la langue. Exemple:

> [équidé] – [manger] – [nourriture] dans lequel chaque élément […] représente une classe de termes. [Nourriture]: avoine, herbe, fourrage, carotte, légumineuses, betteraves, etc…

Ces GC sont nécessairement très nombreux. L'avantage des graphes étant qu'ils peuvent être reliés les uns aux autres de façon diverses ou emboîtés les uns dans les autres. Les problèmes qu'ils posent sont alors de plusieurs natures:

> 1.À quelles conditions un GC peut-il s'attacher avant ou après un autre, ceci dans une régression infinie.
> 2. Quelles sont les éventuelles modifications internes aux différents GC impliquées par leurs agglutinations.

Un générateur de description de plantes virtuelles me servira d'exemple. Ce générateur demande d'abord une modélisation d'un texte de type herbier. Voici deux des modèles possibles :

1. [thl-présente-01]. [thl-découverte-01]. [thl-fleur-01]. [thl-tige-01]. [thl-habitat-01]. [thl-plante-01]. [thl-nom-01].< [thl-proche-01].> [thl-racine-01]. [thl-floraison-01]. [thl-graine-01]. [thl-fruit]. [thl-astre-01]. [thl-leçon-01].
2. [caract10]%%[thl-présente-01]. [thl-découverte-01]. [thl-fleur-01]. [thl-floraison-01]. [thl-tige-01]. [thl-graine-01]. [thl-fruit]. [thl-plante-01]. [thl-racine-01]. [thl-habitat-01]. [thl-nom-01].< [thl-proche-01].> [thl-astre-01]. [thl-leçon-01].

Si l'on élimine les aspects dépendant d'un logiciel de génération donné, on remarque que chacun de ses modèles est constitué d'éléments ordonnés:

[présente-01]	[caract10]	[astre-01]
[découverte-01]	[présente-01]	[caract10]
[fleur-01]	[découverte-01]	[découverte-01]
[tige-01]	[fleur-01]	[fleur-01]
[habitat-01	[floraison-01]	[floraison-01]
[plante-01	[tige-01]	[fruit]
[nom-01]	[graine-01]	[graine-01
[proche-01]	[fruit]	[habitat-01]
[racine-01]	[plante-01]	[leçon-01]
[floraison-01]	[racine-01]	[nom-01]
[graine-01	[habitat-01]	[plante-01
[fruit]	[nom-01]	[présente-01]
[astre-01]	[proche-01]	[proche-01]
[leçon-01]	[astre-01]	[racine-01]
	[leçon-01]	[tige-01]

Ils représentent de légères variantes de structure dans une combinatoire légèrement différente d'ensembles partagés: c'est cette structure (que l'on peut rapidement regrouper en: présentation, description, habitat, plante proche) et le communauté de ses éléments qui va donner la cohérence à l'infinité possible des texte générés comme celui-ci en français puis en espagnol pour montrer que le modèle est translinguistique: Bella Donna sive Linnius Hypericum Digitalis

Certains estiment que cette graminée est la vraie Acarus alba. Plante merveilleuse des arméniens, cette herbacée. La plante la plus proche est la Bryone capricieuse. Les visionnaires lui donnent le nom d'Afghanite absolue. Pousse souvent sur les versants des montagnes orientés à l'ouest. Les fleurs: pétales rougis ou rougeâtres très articulés quoique allongés. "Quand au mois de mai fleurit Acarus alba", ainsi s'exprime le poète Kevin Saunderson dans son volume "le temps de la vie est infime… celui de la mort infini". Dans son fruit, éblouissant, on découvre de petits fruits appelés Morella. Cette plante se divise en de multiples tiges hirsutes et polyédriques portant chacune une fleur. Graines grandes, grises, grisâtres, gravides. La racine: vide, rampante, solide. C'est quand la campagne se dévoile que cet arbuste s'épanouit dans toute sa merveille. Cette spirée apporte la volupté et l'amour dans le couple. Dans son roman "toute parole est infinie" (1428), Susan Howe fait dire à cette plante herbacée: "les choses sont ce qu'elles sont". La première référence aux vertus médicinales de cette plante grimpante se trouve dans l'herbier d'Eduardo Marga de 1815.

> «Silene luminaris sive Muflier de Borgès (Morella graveolens)
> Plantas son necesarias para la supervivencia de los hombres. A esta planta llana pérenne la llaman la "Azalea turbia", planta abundante y olorosa. Planta procedente de Congo Popular Republic, Georges Aperghis la trajo de China en 2001. En la parte superior del tallo aparecen las florcillas tubulosas amarillas y lígulas periféricas blancas. *El tallo fecundo, avaro, malo, curadado. Crece sobre todo en Amazonia.* Las paleobiólogas Marion Xingjian y Myriam Smith tuvieron una larga discusión en que se trataba de saber si esta planta silvestre compuesta es de Integración o de Anterior. Su nombre, “Bulbo generoso”, se lo dio en 1387 Maurice Benayoun. Se puede extraer de su raíz un colorante marrón muy poderoso. Floración de Abril a Octubre. Esta hierba trepadora no tiene mala semilla. Su fruto se puede comer. También cuenta con muy buenas propiedades como aperitivo, pues excita el apetito. Del Investigado del Paraguay se usa la raíz machacada, en las aguas de bebida. En cuanto a sus propiedades, se puede decir que son muy semejantes a las del berro. Morella graveolens florece proclamando: “la crisálida del tiempo engendra a la mariposa del movimiento y de la muerte”.»

Si ce premier niveau de modélisation est indispensable à la perception d’un texte, il n’est cependant pas suffisant. En effet, intervient là ce que l’on peut rapidement appeler le «niveau syntaxique» qui se manifeste sur l’ensemble du texte de façons différentes:

- temps verbaux
- cohérences pronominales
- cohérences sémantiques

Toutes contraintes aboutissant à des descriptions plus « riches » des GC, comme par exemple celle-ci:

> [199#] [32|m_plante] [10|dis-astre]< [122#] [010000000|dis-soigner] [11|dis-maladie]>
> Correspondant au GC élémentaire suivant:
> [plante] --------- [soigner] -------- [maladie]

Mais entrer maintenant dans le détail technique de toutes ces descriptions nous entraînerait trop loin.

Une dernière remarque cependant, un tel GC n’est génératif qu’à la condition que l’un quelconque de ses éléments puisse contenir plus d’un élément. S’il doit aboutir au résultat suivant:

L’Aloe Vera soigne l’eczéma, son utilisation est sans intérêt. Mais, il est bien rare, dans une langue naturelle qu’il en soit ainsi, «soigner», par exemple, en français peut en effet se dire de bien des façons différentes. C’est d’ailleurs ce qui distingue une langue naturelle d’une langue formelle dans laquelle doit entrer le moins d’entropie possible.