Expérimentations de création législative par chaînes de Markov

Définition

On appelle “chaîne de Markov” un processus stochastique consistant en une “marche” aléatoire à travers un graphe, chaque “pas” étant déterminé uniquement par l'étape actuelle (le processus n'a pas de mémoire quant à ce qui s'est produit aux étapes précédentes). Le graphe est en outre pondéré, en cela que les progressions ne sont pas toutes équiprobables.

En appliquant cette idée au langage humain, on peut associer à chaque nœud du graphe un mot, et pondérer les relations entre ces mots en fonction de la probabilité, dans un corpus donné, que ces mots se succèdent directement. Par exemple, depuis le nœud “Tribunal”, et en utilisant pour corpus l'ensemble du droit fédéral suisse en vigueur, la probabilité est plutôt grande que le prochain mot soit “fédéral” (742 occurrences, pour 3277 occurrences du mot “Tribunal”, soit 22.6%) ; au contraire, il est très rare que le mot suivant soit “uranais” (une seule occurrence sur 3277). Évidemment, l'immense majorité des mots du corpus ne s'observent *jamais* juste après le mot “Tribunal”.

Ainsi, lors de la génération d'une phrase nouvelle, et en admettant que le dernier mot choisi fut “Tribunal”, les chances sont plutôt élevées que le mot suivant soit “fédéral”. À l'étape suivante, c'est le mot “fédéral” qui servira de base ; le mot d'après sera choisi selon la même procédure. Ainsi, de proche en proche, on peut reconstruire une phrase qui pourra donner l'impression d'être issue du corpus choisi.

Limites

Nous l'avons dit, le processus n'a pas de mémoire à proprement parler. Ainsi, le seul contexte à disposition pour progresser est le dernier mot généré. Il est cependant possible d'étendre le procédé en utilisant comme unité d'information les paires de mots, au lieu des mots eux-mêmes. Il est à noter que procéder ainsi augmente rapidement la complexité de l'algorithme et nécessite une importante quantité de mémoire. Le corpus susmentionné (l'ensemble du droit fédéral suisse en vigueur) compte en effet plus de 48'000 mots distincts, mais plus de 380'000 paires distinctes de deux mots successifs séparés par un espace, et plus de 740'000 “triplets” distincts.

En outre, même en procédant ainsi, on ne règle pas (ou très partiellement seulement) le problème de la cohérence grammaticale des phrases générées. Ce problème peut être mitigé en passant les phrases générées au crible d'un vérificateur grammatical automatique et en ne conservant que celles qui ne contiennent aucune erreur ; néanmoins, même ces phrases-ci peuvent apparaître comme très peu naturelles. Ainsi, la phrase suivante a été considérée comme grammaticalement correcte par deux logiciels distincts :

`Le candidat qui se trouve à l'étranger par un service militaire n'est pas possible.`

"Se trouver à l'étranger **par** un service militaire" est grammaticalement correct (on s'en convainc par le fait que "pour" aurait du sens, et que "par" et "pour" sont tous deux de la même catégorie grammaticale, à savoir des prépositions), mais cette proposition n'a pourtant aucun sens. En outre, on ne dira jamais d'un candidat qu'il n'est "pas possible" (du moins pas dans un contexte juridique). Ces subtilités échappent aux correcteurs grammaticaux.

Intérêt

Malgré ses limites, une telle approche peut faciliter le travail de saisie de contenu textuel juridique par un être humain, par exemple un greffier au sein d'une juridiction, en suggérant automatiquement les mots les plus probables compte tenu de ce qui vient d'être entré. Outre le gain de temps et d'énergie, cela permet sans doute, dans une certaine mesure, de contribuer au respect de la précision terminologique qui sert la sécurité juridique et l'accessibilité du droit (Refaire la loi, ch. 5.3.3.2.a.III).

Exemples

Exemples réalistes

Ces exemples sont non seulement considérés comme grammaticalement corrects par les correcteurs automatiques, mais ils semblent en outre (plus ou moins) naturels à l'auteur.

- `Sur demande du requérant, elle lui communique immédiatement les autres réserves.`
- `Sont réputés équivalents d'autres temps de présence ou du capital social d'une société en nom collectif ou en anglais.`
- `Elle peut effectuer des contrôles de sécurité en collaboration avec l'Agence.`
- `Elle est régie par le preneur de crédit avec une ferrure en état.`
- `Le Conseil fédéral édicte les directives conjointement avec des instituts financiers.`
- `Le rapport décrit les méthodes et critères d'évaluation nécessaires.`
- `La pente ne doit pas être interrompue.`
- `Le système est utilisé par les cantons pour l'intégration.`
- `Le Tribunal fédéral connaît des recours contre les fonctionnaires et autorités du canton de Bâle-Ville.`
- `La poursuite pénale doit être portée sans délai par des feux blancs.`

Exemples moins réalistes

Ces exemples sont considérés comme grammaticalement corrects par les correcteurs automatiques (ce qui indique les limites de tels correcteurs, du moins ceux utilisés par l'auteur).

- `Toute clause contenant des données du système de référence prend en charge s'avère nécessaire, la personne concernée a déposé la proposition, la période de contrôle et de concession correspond au prix de l'offre de boissons spiritueuses est limitée à un alias est réceptionné dans la fourniture d'un service de protection en millimètres doit être réalisé à l'extérieur ainsi que les résultats des contrôles et aux autorités compétentes en matière d'hygiène, de sécurité des données.`
- `Les cantons indiquent à l'autorité locale compétente.`
- `Il assure la bonne foi et de maisons d'habitation comptant moins de 40 francs par personne interposée, livre, transmet ou non remboursés à la partie requérante, ordonner les mesures appropriées doivent permettre de constater qu'il est remis au marin.`
- `L'obligation de témoigner au sens de la levée de la lecture.`
- `Il rend tous les deux jours au moins.`

Le dernier exemple pourrait à la limite se trouver dans une liste de critères alternatifs permettant de considérer, en l'absence d'un certificat médical, qu'un travailleur est manifestement malade.