Modèles psycholinguistiques pour la segmentation en mots

Alexis NASR Arnaud REY

October 2022

La segmentation en mots est la tâche qui consiste à segmenter un signal acoustique en segments correspondant à des mots. Il s'agit d'une tâche complexe qui suppose l'utilisation d'un grand nombre d'indices, acoustiques, lexicaux, syntaxiques, sémantiques ...

[Saffran et al., 1996] a montré à l'aide d'expériences psycholinguistiques simples que l'être humain était capable d'effectuer une segmentation en mots à partir d'un signal beaucoup plus pauvre, en utilisant simplement des régularités statistiques. Ces expériences ont porté sur un signal acoustique composé d'une concaténation de 6 mots construits à partir d'un inventaire de douze syllabes. Le signal était produit par un synthétiseur de parole et la parole produite ne comportait aucun indice prosodique (pas de pauses, pas d'accent ...).

[Perruchet and Vinter, 1998] a proposé un algorithme simple, appelé PARSER, de segmentation en mots qui permet de reproduire une partie des comportements observés par [Saffran et al., 1996] sur des être humains. Le dispositif expérimental n'est pas tout à fait le même que dans [Saffran et al., 1996] dans la mesure où, pour des raisons de simplicité, il ne s'agit pas d'un signal acoustique, mais d'une chaîne de caractères. PARSER tire profit de régularités statistiques contenues dans le signal sans calculer explicitement de probabilités. Le modèle repose sur un jeu de 6 hyperparamètres dont les valeurs ont une influence déterminante sur les performances du modèle.

L'objet du stage est de partir du modèle proposé par [Perruchet and Vinter, 1998] et de le faire évoluer.

La première évolution consiste à utiliser des données plus réalistes, en particulier un inventaire de syllabes et un lexique plus riches. L'objet de cette partie est d'étudier l'évolution des performances de PARSER au fur et à mesure que les données deviennent plus complexe. On utilisera deux types de données. D'une part, des données synthétiques comme celles utilisées par [Saffran et al., 1996], en augmentant le nombre de syllabes mais aussi la longueur et le nombre de mots différents. D'autre part des données plus naturelles, avec des syllabes et des mots du français.

La seconde évolution concerne la valeur des paramètres. Chacun de ces paramètres joue un rôle important dans le comportement de PARSER. On étudiera l'influence de chacun de ces paramètres sur les performances de l'algorithme,

notamment lorsque les données deviennent plus complexe. On étudiera aussi la possibilité que le système optimise lui même la valeur des paramètres.

La troisième évolution consiste à placer PARSER dans des conditions plus proches de celle de l'expérience originelle de [Saffran et al., 1996] en prenant en entrée non plus une chaîne de caractère, mais un signal acoustique.

Ce stage s'inscrit dans une perspective plus générale [Goyal and Bengio, 2020, Perruchet and Vinter, 2021] qui consiste à étudier dans quelle mesure les obstacles auxquels sont confrontés les systèmes actuels de TAL fondés sur l'apprentissage profond peuvent être dépassés en introduisant des biais inductifs issus de la psycholinguistique, dont PARSER est un exemple.

References

[Goyal and Bengio, 2020] Goyal, A. and Bengio, Y. (2020). Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*.

[Perruchet and Vinter, 1998] Perruchet, P. and Vinter, A. (1998). Parser: A model for word segmentation. *Journal of memory and language*, 39(2):246–263.

[Perruchet and Vinter, 2021] Perruchet, P. and Vinter, A. (2021). The self-organizing con-sciousness: Implications for deep learning. *Trends Artif Intell*, 5(1):87–94.

[Saffran et al., 1996] Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.