Segmenteur morphologique pour le français

Encadrement : beatrice.daille@univ-nantes.fr et ygor.gallina@univ-nantes.fr
Lieu de TER : LS2N (Site de la FST)
Nombre d'étudiant·es : 1/2 avec une appétence pour le TAL et la linguistique

Contexte

La segmentation d'un texte en tokens (unités textuelles, par exemple des mots) est la première étape de tout travail en traitement automatique des langues. Une technique simple consiste a considérer le caractère " " comme délimiteur de token. La technique la plus utilisée aujourd'hui consiste a segmenter un texte en « sous-mots », ces sous-mots correspondent à des séquences de caractères fréquentes dans un corpus. Ils sont calculés automatiquement grâce à l'algorithme BPE[1] basé sur la fréquence.

Malgré ses avantages cette segmentation automatique produit un découpage qui ne respecte pas la morphologie des mots. Par exemple le tokeniseur du modèle camembert-base[1] segmente `"décapuchonnement"` en `"déc_ap_uch_onnement"`, une segmentation morphologique donnerait `"dé_capuchonne_ment"`, où `"dé"` est le préfixe, `"capuchonne"` la racine et `"ment"` le suffixe.

Sujet

L'objectif de ce travail est de créer un segmenteur morphologique pour le français, pour cela nous nous baserons sur des ressources linguistiques existante (cf. références).
- Identifier les ressources linguistiques existantes pour faire de l'analyse morphologique
- Proposer un segmenteur morphologique pour le français
- Evaluer le segmenteur morphologique par rapport à d'autres stratégies de segmentation dans une tâche de TAL

Références

[1] Gage, P. (1994). A new algorithm for data compression. C Users Journal, 12(2), 23-38. pennelynn.com/Documents/CUJ/HTML/94HTML/19940045.HTM

[2] Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia - MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology aclanthology.org/2021.sigmorphon-1.5 github.com/kbatsuren/MorphyNet

[3] Hathout, N., Sajous, F., Calderone, B., Namer, F. (2020). Glawinette: a linguistically motivated derivational description of French acquired from GLAWI. In Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020), pp. 3870-3878, Marseille, 2020. redac.univ-tlse2.fr/lexiques/glawinette/Hathout-2020-LREC-Glawinette.pdf redac.univ-tlse2.fr/lexiques/glawinette.html

[4] McCarthy, A.D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S.J., Nicolai, G., Silfverberg, M. and Arkhangelskij, T., (2020). UniMorph 3.0: Universal Morphology.. Proceedings of LREC 2020. aclanthology.org/2020.lrec-1.483 github.com/unimorph/fra

1 https://huggingface.co/camembert-base