Sujet de Stage de M2

Titre : Evaluation du volume des boules de mots

Encadrant

Jean-Christophe Janodet, Labo IBISC, Univ. Evry, Université Paris-Saclay
jeanchristophe.janodet@univ-evry.fr

Contexte

Une *boule de mots* est un langage fini, défini sur un alphabet fixé A. C'est plus précisément l'ensemble de tous les mots $w \in A^*$ qui sont à une distance au plus $r \geq 0$ d'un mot $c \in A^*$ donné, appelé le *centre* de la boule : $B(c, r) = \{w \in A^* : d(c, w) \leq r\}$.

De nombreuses distances peuvent être considérées : distance de Hamming, distance de Levenshtein, distance d'édition. Toutes ces métriques sont à la base de nombreux travaux en bioinformatique [7, 6], en modélisation de la langue [2, 1], ou en reconnaissance des formes [10, 4].

D'un point de vue pratique, on retrouve ces boules dans plusieurs travaux (sans qu'elles soient toujours nommées). Ainsi, dans le cadre de la recherche de motifs approximatifs, on peut être amené à chercher toutes les chaînes proches d'une certaine chaîne cible [4], ou au contraire, à retrouver une chaîne cible à partir de versions dégradées de celle-ci [8].

De même, la tâche d'un correcteur orthographique peut être décrite comme une recherche dans l'intersection de deux langages, le dictionnaire lui-même, et une boule autour du mot à corriger [11]. En Inférence Grammaticale, elle apparaissent lorsqu'on cherche à apprendre des grammaires à partir de données bruitées [12]. Dans [3, 5], la question de leur apprenabilité a été étudiée de façon systématique, dans le cadre des paradigmes d'apprentissage standard.

Problème, résultats attendus

De nombreuses questions, simples dans leur énoncé, restent ouvertes concernant les boules de mots. C'est en particulier le cas du *volume* d'une boule, c'est-à-dire du nombre de mots qu'elle contient, en fonction de la taille du centre, du rayon, et de la distance utilisée.

Dans [9], les auteurs proposent d'aborder cette question d'un point de vue expérimental, à l'aide d'un algorithme probabiliste fournissant des estimations du volume des boules. Toutefois, cette proposition est à l'état d'ébauche : l'algorithme n'est pas implémentable en l'état, et l'étude expérimentale elle-même est à peine esquissée.

L'objectif principal du travail est de proposer un nouvel algorithme, corrigeant et améliorant la version précédente, et qui soit exploitable pour mener une étude expérimentale à une échelle raisonnable.

Une seconde question pourra ensuite être étudiée, indépendamment de la première : celle de la taille d'un automate minimal qui reconnaît une boule de mots.

References

[1] J.-C. Amengual and P. Dupont. Smoothing probabilistic automata: An error-correcting approach. In *Proc. ICGI'00*, pages 51–64. LNAI 1891, 2000.

[2] J.-C. Amengual, A. Sanchis, E. Vidal, and J.-M. Benedí. Language simplification through error-correcting and grammatical inference techniques. *Machine Learning Journal*, 44(1-2):143–159, 2001.

[3] L. Becerra-Bonache, C. de la Higuera, J.-C. Janodet, and F. Tantini. Learning balls of strings from edit corrections. *Journal of Machine Learning Research*, 9:1823–1852, 2008.

[4] E. Chávez, G. Navarro, R. A. Baeza-Yates, and J. L. Marroquín. Searching in metric spaces. *ACM Computing Surveys*, 33(3):273–321, 2001.

[5] C. de la Higuera, J.-C. Janodet, and F. Tantini. Learning languages from bounded resources: The case of the dfa and the balls of strings. In *Proc. ICGI'08*, pages 43–56. LNAI 5278, 2008.

[6] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, 1998.

[7] D. Gusfield. *Algorithms on Strings, Trees, and Sequences - Computer Science and Computational Biology*. Cambridge University Press, 1997.

[8] T. Kohonen. Median strings. *Pattern Recognition Letters*, 3:309–313, 1985.

[9] H. Koyano and M. Hayashida. Volume formula and growth rates of the balls of strings under the edit distances. 2022.

[10] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[11] K. U. Schulz and S. Mihov. Fast string correction with Levenshtein automata. *Int. Journal on Document Analysis and Recognition*, 5(1):67–85, 2002.

[12] F. Tantini. *Inférence grammaticale en situations bruitées*. PhD thesis, University of Saint-Etienne, 2009.