Stage M2R CEDRIC
Réécriture sémantique de requêtes et recherches avancées sur un corpus hiéroglyphique

Serge Rosmorduc (MIM), Nicolas Travers (Vertigo)

Mots-clés : Hiéroglyphes, alignement, profilage, word2vec,

Contexte

Ce stage fait suite à un travail précédent sur la production d'un moteur de recherche basé sur notre corpus hiéroglyphique *Ramsès* [1,2]. Le modèle complexe intègre les données selon plusieurs dimensions (graphies, phonèmes, traductions, phrases, métadonnées…) et permet d'effectuer des requêtes riches sur le corpus dans le domaine de la recherche d'informations [3,4]. La communauté d'égyptologues, l'Université de Lièges notamment, y voit un intérêt en termes d'évolution des méthodologies de travail en vue de partager et rechercher les connaissances. Dans cette optique, ce premier travail sera prochainement soumis à la communauté *humanité numérique* pour valider notre approche, puis étendu/amélioré avec les éventuelles avancées de ce nouveau stage.

Sujet de stage

Dans la continuité, trois pistes d'améliorations sont envisagées afin d'améliorer à la fois la pertinence des résultats que les possibilités d'interrogation :

- Le corpus ayant été produit au cours du temps par différents égyptologues, les habitudes de traductions / choix des graphies / phonèmes ne sont pas uniformes. De fait, les résultats des recherches ne sont pas toujours pertinents de manière sémantique. Nous envisageons d'effectuer une analyse automatique de chaque dimension des données avec *word2vec* (similaire à [5]) pour extraire des distances sémantiques dans les usages des égyptologues. Les vecteurs résultats seront alors exploités pour réécrire les requêtes afin d'élargir l'espace de recherche, tout en pondérant ces nouveaux résultats par la distance sémantique entre les vecteurs requêtes.

  De plus nous envisageons d'étendre cette approche pour les égyptologues à des fins d'homogénéisation de corpus en analysant les tendances et profilage des traductions.

- Actuellement, le langage de requêtes permet de rechercher dans les différentes dimensions de traduction en même temps et de classer le résultat par pertinence, tout en gardant les structures de phrases. Nous envisageons d'étendre les possibilités du langage avec des notions d'alignement de traductions. En effet la même donnée est présente sous trois formes dans le document, et pouvoir aligner

les variantes de traductions dans la requête permettrait d'effectuer des recherches plus riches et proche des besoins de l'égyptologue (graphie traduite d'une certaine manière, corrélation de graphies, évolutions de traductions). Par ailleurs, l'alignement permettra d'améliorer la pertinence des résultats puisqu'il permettra d'ajouter un calcul de distance de traduction et ainsi d'améliorer les scores obtenus pour chaque requête. Cette notion d'alignement de « vues » sur les données n'est pas présente dans la littérature en Recherche d'Information car les données produites sous différentes versions sont rarement corrélées (en dehors de l'identifiant de document).

- D'autre part, l'encodage des graphies intègre une hiérarchie de caractéristiques pour chaque hiéroglyphe. Il serait intéressant d'améliorer la recherche des documents via les graphies avec cette notion hiérarchique. Nous pourrions comparer cela à des recherches d'emoji en Unicode 6.0 ayant vu le jour sous elasticsearch récemment (JoliCode [6])

Ce stage reposera donc sur les résultats obtenus pour améliorer le moteur de recherche afin de l'intégrer à Ramsès et effectuer une évaluation de la qualité des résultats en partenariat avec des égyptologues.

Références :

[1] S. Polis, S. Rosmorduc, J. Winand. "Ramses goes online. An annotated corpus of Late Egyptian texts in interaction with the Egyptological community", International Congress of Egyptologists, August 2015, Vol. XI, pp.n/a, Florence, Italie,

[2] S. Rosmorduc, "Automated Transliteration of Egyptian Hieroglyphs" in N. Strudwick ed., *Information Technology and Egyptology in 2008,* p. 167-183.

[3] N. Travers. "Putting into Practice: Full-Text Indexing with LUCENE", Titre du livre: "*Web Data Management*", February 2012, Cambridge University Press, pp. 355--363, (isbn: 9781107012431)

[4] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008

[5] Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana, « *Improving Document Ranking with Dual Word Embeddings* ». WWW'16

[6] D. Alexandre. « Search for Emoji with Elasticsearch » https://jolicode.com/blog/search-for-emoji-with-elasticsearch. Décembre 2016