Résumés de thèses

Rubrique préparée par Sylvain Pogodalla

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
sylvain.pogodalla@inria.fr

Pierre GODARD : pierre@lpdi.org

Titre : Découverte non supervisée de mots pour outiller la linguistique de terrain

Mots-clés : apprentissage non supervisé, segmentation automatique en mots, alignement bilingue, modèles bayésiens, langues peu dotées.

Title: *Unsupervised Word Discovery for Computational Language Documentation*

Keywords: *unsupervised learning, automatic word segmentation, bilingual alignment, Bayesian models, low-resource languages.*

Thèse de doctorat en informatique, école doctorale *sciences et technologies de l'information et de la communication*, LIMSI-CNRS, Université Paris-Saclay, sous la direction de François Yvon (Pr, Université Paris-Sud, LIMSI-CNRS) et Laurent Besacier (Pr, Université Grenoble-Alpes, LIG). Thèse soutenue le 16/04/2019.

Jury : M. François Yvon (Pr, Université Paris-Sud, LIMSI-CNRS, codirecteur), M. Laurent Besacier (Pr, Université Grenoble-Alpes, LIG, codirecteur), M. Pierre Zweigenbaum (DR, CNRS, LIMSI, président), M. Christophe Cerisara (CR, CNRS, LORIA, rapporteur), M. Adam Lopez (Associate professor, University of Edinburgh, ILCC, rapporteur), M. Emmanuel Dupoux (Directeur d'études, EHESS, LSCP, examinateur).

Résumé : *La diversité linguistique est actuellement menacée : la moitié des langues connues dans le monde pourraient disparaître d'ici la fin du siècle. Cette prise de conscience a inspiré de nombreuses initiatives dans le domaine de la linguistique documentaire au cours des deux dernières décennies, et 2019 a été proclamée* année internationale des langues autochtones *par les Nations Unies, pour sensibiliser le public à cette question et encourager les initiatives de documentation et de préservation.*

Néanmoins, ce travail est coûteux en temps et le nombre de linguistes de terrain, limité.

Par conséquent, le domaine émergent de la documentation linguistique computationnelle *vise à favoriser le travail des linguistes à l'aide d'outils de traitement automatique. Le projet* Breaking the Unwritten Language Barrier (BULB)*, par exemple, constitue l'un des efforts qui définissent ce nouveau domaine et réunit des linguistes et des informaticiens. Cette thèse examine le problème particulier de la découverte de mots dans un flot non segmenté de caractères, ou de phonèmes, transcrits à partir du signal de parole dans un contexte de langues très peu dotées. Il s'agit principalement d'une procédure de segmentation qui peut également être couplée à une procédure d'alignement lorsqu'une traduction est disponible.*

En utilisant deux corpus en langues bantoues correspondant à un scénario réaliste pour la linguistique documentaire, l'un en mboshi (République du Congo) et l'autre en myene (Gabon), nous comparons diverses méthodes monolingues et bilingues de découverte de mots sans supervision. Nous montrons ensuite que l'utilisation de connaissances linguistiques expertes au sein du formalisme des Adaptor Grammars *peut grandement améliorer les résultats de la segmentation, et nous indiquons également des façons d'utiliser ce formalisme comme outil de décision pour le linguiste. Nous proposons aussi une variante tonale pour un algorithme de segmentation bayésien non paramétrique qui utilise un schéma de repli modifié pour capturer la structure tonale. Enfin, pour tirer parti de la supervision faible d'une traduction, nous proposons et étendons une méthode de segmentation neuronale basée sur l'attention et améliorons significativement la performance d'une méthode bilingue existante.*

URL où le mémoire peut être téléchargé :
http://www.theses.fr/s156321

Maxime WARNIER : maximewarnier@gmail.com

Titre : Contribution de la linguistique de corpus à la constitution de langues contrôlées pour la rédaction technique : l'exemple des exigences de projets spatiaux

Mots-clés : exigences, spécifications, langue contrôlée, genre textuel, corpus.

Title: *A Methodology for Creating Controlled Natural Languages for Technical Writing Based on Corpus Analysis: a Case Study on Requirements Written for Space Projects*

Keywords: *requirements, specifications, controlled language, textual genre, corpus.*

Thèse de doctorat en sciences du langage, CLLE CNRS, Université Toulouse - Jean Jaurès, sous la direction de Anne Condamines (DR, CNRS). Thèse soutenue le 10/09/2018.

Jury : Mme Anne Condamines (DR, CNRS, directrice), M. Thierry Charnois (Pr, Université Paris 13, rapporteur), Mme Natalie Kübler (Pr, Université Paris Diderot, présidente), M. Ulrich Heid (Pr, Universität Hildesheim, Allemagne, examinateur), M. Ludovic Tanguy (MC, Université Toulouse - Jean Jaurès, examinateur).

Résumé : *L'objectif de notre travail, qui émane d'une demande de la sous-direction Assurance Qualité du CNES (Centre National d'Études Spatiales), est d'augmenter la clarté et la précision des spécifications techniques rédigées par les ingénieurs préalablement à la réalisation de systèmes spatiaux. L'importance des spécifications (et en particulier des exigences qui les composent) pour la réussite des projets de grande envergure est en effet désormais très largement reconnue, de même que les principaux problèmes liés à l'utilisation de la langue naturelle (ambiguïtés, flou, incomplétude) sont bien identifiés. Dès lors, de nombreuses solutions, plus ou moins formalisées, ont été proposées et développées pour limiter les risques d'interprétation erronée — dont les conséquences potentielles peuvent se révéler extrêmement coûteuses — lors de la rédaction des exigences, allant des langages logiques aux guides de rédaction, en passant par des outils de vérification semi-automatique.*

Nous pensons que pour qu'elle soit réellement adoptée par les ingénieurs du CNES (qui ne sont actuellement pas tenus de suivre des règles de rédaction), la solution que nous nous efforçons de mettre au point se doit d'être à la fois efficace (autrement dit, elle doit limiter sensiblement le risque langagier) et aisée à mettre en place (autrement dit, elle ne doit pas bouleverser trop profondément leurs habitudes de travail, ce qui la rendrait contre-productive). Une langue contrôlée (en anglais : controlled natural language*), c'est-à-dire un ensemble de règles linguistiques portant sur le vocabulaire, la syntaxe et la sémantique, nous paraît être une réponse idéale à ce double besoin — pour autant qu'elle reste suffisamment proche de la langue naturelle. Or, les langues contrôlées pour la rédaction technique déjà existantes que nous avons envisagées, bien qu'élaborées par des experts du domaine, ne nous semblent pas toujours pertinentes d'un point de vue linguistique : certaines règles sont trop contraignantes, certaines ne le sont pas assez, d'autres encore ne se justifient pas vraiment.*

Nous voudrions donc définir une langue contrôlée pour la rédaction des exigences en français au CNES. L'originalité de notre démarche consiste à systématiquement vérifier nos hypothèses sur un corpus d'exigences (constitué à partir d'authentiques spécifications de projets spatiaux) à l'aide de techniques et d'outils de traitement automatique du langage existants, dans l'optique de proposer un ensemble cohérent de règles (nouvelles ou inspirées de règles plus anciennes) qui puissent ainsi être vérifiées semi-automatiquement lors de l'étape de spécification et qui, surtout, soient conformes aux pratiques de rédaction des ingénieurs du CNES. Pour cela, nous nous appuyons notamment sur l'hypothèse de l'existence d'un genre textuel, que nous tentons de prouver par une analyse quantitative, ainsi que sur les notions de normalisation et normaison. Notre méthodologie combine les approches corpus-based *et* corpus-driven *en tenant compte à la fois des règles imposées par deux autres langues*

contrôlées (dont l'adéquation avec des données réelles est discutée au travers d'une analyse plus qualitative) et des résultats offerts par des outils de fouille de textes.

URL où le mémoire peut être téléchargé :
https://tel.archives-ouvertes.fr/tel-02062833