Retour de mission LREC 2014 Reykjavik
Antonio Balvet

The 9th edition of the Language Resources and Evaluation Conference, 26-31 May, Reykjavik

Au cours de l'édition 2014 de la conférence LREC se sont dessinées les tendances de fond du domaine des Technologies de la Langue pour les années à venir (Horizon2020):
- fédération de ressources et d'outils pour le traitement de la langue, sous l'impulsion du groupe d'intérêt META-NET. La plate-forme web distribuée Meta-share, notamment, est appelée à jouer un rôle central dans le recensement et la diffusion de ressources et d'outils pour le TAL, quel que soit le régime de licence ou le mode de diffusion.
- standardisation des *worklows*, par le biais de plate-formes open-source paramétrables: la création de processus de traitement est facilitée grâce à l'adoption d'une logique modulaire, ainsi qu'à la standardisation des différentes étapes de traitement des langues. Notamment, les étapes de segmentation en unités linguistiques de base (mots, phrases), étiquetage en parties du discours et analyses syntaxiques de surface sont fournies "clés en main" pour un nombre croissant de langues. Ainsi, par exemple, la plate-forme OpenER devrait être disponible dans les mois prochains. Elle est issue d'un projet financé dans le cadre du FP7 de l'UE, réunissant entreprises privées et laboratoires académiques. Elle est orientée vers le suivi d'opinion et de sentiments en 6 langues, mais d'autres langues pourront être ajoutées. La mise à disposition de telles plate formes est conforme à l'orientation annoncée lors du forum HLT 2013 à Bruxelles par la représentante de la branche "projets de recherche" de l'UE : le développement des partenariats public-privé via des plate formes et standards ouverts, des investissements structurels en faveurs des technologies de la langue au service de la citoyenneté européenne et de l'économie numérique. Ces plate formes permettront, par exemple, aux entreprises partenaires de STL d'adopter à moindre coût de bonnes pratiques en matière de traitement de la langue.
- les Linguistic Linked Data: exploitation de ressources lexicales structurées en adoptant les standards du moment (LMF, RDF), et connectées, disponibles sous la forme de web-services.
- la montée en puissance du *crowdsourcing* (annotation par les foules, via des protocoles commerciaux de type "*mechanical Turk*" d'Amazon, avec rémunération à la tâche) et autres *serious games*, pour l'élaboration de ressources linguistiques ou l'annotation de corpus à moindre coûts. Toutefois, les retours de différentes expériences de *crowdsourcing* sont assez mitigés : dans la plupart des cas, les annotations linguistiques (sémantiques par ex.) sont d'un niveau de complexité trop élevé pour être confiées à des non-spécialistes, à moins de réduire considérablement la difficulté, auquel cas les annotations ainsi récoltées s'avèrent relativement peu intéressantes et exploitables. Les *serious games* semblent présenter un potentiel plus intéressant, notamment pour l'enrichissement de ressources lexicales (ex: ajout de relations sémantiques par des utilisateurs/joueurs). Toutefois, cette stratégie de constitution de ressources lexicales à moindres coûts demande en amont un gros travail de réflexion pour transformer une tâche linguistique en jeu motivant pour des utilisateurs non-spécialistes, ainsi qu'un investissement important en termes de développement d'applications mobiles (les *serious games* sont typiquement proposés comme applications pour téléphones mobiles type iPhone) et de gestion de la communauté des utilisateurs/joueurs (stratégie de motivation des bons joueurs via récompenses, neutralisation des joueurs peu actifs ou trolls). En définitive, ce type de stratégie semble envisageable dans des domaines restreints (lexiques génériques à couverture limitée), mais semble peu applicable pour des domaines très spécialisés (ex: médecine, finance etc.).

- la constitution de ressources linguistiques pour des langues peu dotées par "ricochet", à partir de ressources linguistiques de langues typologiquement proches. Par ex. : induction de règles d'analyse syntaxique pour le serbe à partir de corpus annotés en tchèque, polonais et croate.

Enfin, la traduction automatique statistique, à partir de corpus alignés ou simplement comparables, reste un thème de prédilection pour la communauté LREC, ce qui est en droite ligne avec les perspectives du programme Horizon2020.