Thèse IRISA / ANR TREMoLo

Caractérisation de registres de langue par extraction de motifs séquentiels

Contexte

Les registres de langue, c'est-à-dire l'utilisation de mots et formes syntaxiques dédiées à un auditoire ou contexte particulier, sont un trait bien connu de la langue. Ces registres ont une forte influence sur l'expressivité véhiculée par un énoncé, par exemple un registre familier peut dénoter une conversation entre amis, un milieu social peu favorisé... Extraire et manipuler cette dimension expressive des textes est un sujet encore peu étudié en traitement automatique des langues (TAL).

Le sujet de thèse étudie l'emploi des registres de langue et vise à développer des méthodes automatiques de caractérisation de textes d'un registre par rapport à un autre. Pour cela, le sujet propose de s'appuyer sur l'extraction de motifs langagiers spécifiques à des registres donnés, c'est-à-dire les régularités d'ordre morphologiques, syntaxiques et lexicales propres à des textes de ces registres, et sur le regroupement de ces motifs en grandes familles. La langue privilégiée dans le projet sera le français. Les registres de langue font déjà l'objet de travaux en linguistique, ce qui offre un cadre scientifique solide au démarrage de la thèse.

Cette thèse se situe dans le cadre du projet ANR TREMoLo dédié à la transformation automatique de registre. La thèse est donc une étape importante de recherche exploratoire vers des applications à plus long terme sur l'adaptation et la modulation des interactions humain-machine, c'est-à-dire la possibilité d'appliquer le style, l'attitude, ou encore l'émotion d'un ensemble de textes ou phrases de référence à d'autres initialement d'un type différent. À terme, les méthodes développées pourraient par exemple permettre de confronter des discours oraux spontanés et préparés, des articles de vulgarisation à des documents techniques ou encore des articles de presse avec des messages Twitter traitant de la même actualité.

Objectifs

En considérant deux registres A et B (par exemple, soutenu et familier) pour lesquels des corpus textuels sont disponibles mais dont les thèmes abordés ou structurations ne sont pas les mêmes, les objectifs de la thèse se décomposent en trois tâches principales :

1. **Décrire les textes : caractériser des registres de langue en termes de marqueurs linguistiques et produire ces caractéristiques (ou descripteurs) pour les corpus étudiés.** D'une part, le but est de conforter (ou non) des marqueurs considérés comme typiques de certains registres de langue et de découvrir de nouveaux patrons représentatifs de ces registres. D'autre part, cette tâche vise à mettre en place une chaîne d'annotation automatique à la fiabilité évaluée.
2. **Extraire des motifs langagiers discriminants : proposer de nouveaux algorithmes d'extraction de motifs séquentiels émergents, en considérant le contexte de ces derniers.** Cela implique à la fois la production de nouveaux algorithmes par rapport à l'état de l'art en fouille de données ainsi que le développement de démarche de validation des motifs produits, notamment au regard de leurs intelligibilité et significativité pour l'utilisateur.

3. **Regrouper et mettre en correspondance les motifs : clusteriser et fouiller dans les motifs discriminants afin de produire des groupes de motifs similaires et des règles de transformations.** Cette tâche vise à poser de premiers jalons pour moduler des textes ainsi qu'à confronter les connaissances linguistiques du domaine avec celles apprises automatiquement.

Localisation

Le doctorat sera effectué dans les locaux de l'IRISA, campus de Vannes. Des déplacements à Lannion et Paris seront également possibles au cours de la thèse en raison du co-encadrement multi-sites.

Formation et compétences

- Niveau master 2 ou école d'ingénieur en informatique ou TAL requis ;
- Des connaissances en TAL et/ou de la fouille de données et/ou techniques d'apprentissage automatique ;
- Un très bon niveau en français et un bon niveau en anglais (écrit et oral) ;
- De bonnes compétences en programmation (Perl, Python, C++…) ;
- Une bonne connaissance de l'environnement de travail Unix ;
- De bonnes qualités rédactionnelles (précision, clarté).

Début de la thèse

1er octobre 2017.

Candidature

Celle-ci comprendra les éléments suivants (éléments obligatoires en gras): **CV détaillé, lettre de motivation**, **relevés de notes** (avec le classement si possible), **contacts pour recommandation**, rapport(s) de stage(s).

Avant le mardi 2 mai 2017.

Candidature à adresser à toutes les adresses emails suivantes :

del.battistelli@gmail.com, nicolas.bechet@irisa.fr et gwenole.lecorve@irisa.fr.

Références bibliographiques

[1] Gadet, F. (1996). Niveaux de langue et variation intrinsèque. Bensimon & Coupaye, eds, pp. 17-40.

[2] Quiniou S., Cellier P., Charnois T., et Legallois D. (2012). *What about Sequential Data Mining Techniques to Identify Linguistic Patterns for Stylistics?*, in Proceedings of CICLing,, pp. 166-177.

[3] Stamatatos, E. (2009). *A survey of modern authorship attribution methods*. Journal of the American Society for information Science and Technology, 60(3), pp. 538-556.

[4] Neubig, G., Mori, S., et Kawahara, T. (2009). *A WFST-based log-linear framework for speaking-style transformation*. In Proceedings of Interspeech, pp. 1495-1498.