Informatique pour le traitement automatique des langues

Mots clés

informatique, entité nommée spatiale, nom de lieu, apprentissage automatique, similarité, distance lexicale, TEI, XML, traitement automatique des langues

Contexte

Ce projet s'intègre au projet Matriciel : "Lieux des migrants à travers des récits de vie : perceptions, émotions, mots, cartes" (projet PEPS UPE/CNRS 2016). Le corpus de travail est constitué de récits de vie de migrants transcrits et un des objectifs du projet est de fournir des outils automatiques pour aider à leur analyse, cette analyse étant guidée par l'identification automatique des noms de lieux et des sentiments associés à ces lieux.

Les résultats d'analyse sont restitués dans un format cartographique qui permet de présenter sous forme synoptique les lieux et sentiments évoqués dans les différents récits.

Les domaines abordés dans ce projet sont la cartographie et le "traitement automatique des langues" (TAL) ; les données manipulées sont essentiellement des textes et des informations lexicales, syntaxiques, sémantiques concernant les mots des textes, ces informations étant le plus souvent écrites sous forme d'annotations. Les questions de recherche explorées dans ce projet nécessitent des traitements informatiques spécifiques, écrits en JAVA dans le contexte d'utilisation de GATE[1].

GATE est une interface d'intégration de traitements informatiques appliqués à des textes dans laquelle une chaîne de modules statistiques et linguistiques a été développée[2]. Pour le moment, cette chaîne annote automatiquement, dans un texte :

- les mots de sentiments à l'aide de dictionnaires de sentiments et du plugin ANNIE de GATE ;
- les noms propres de lieux à l'aide de dictionnaires de noms propres de lieux et du plugin ANNIE de GATE ;
- des noms communs de lieu à l'aide d'un modèle d'apprentissage automatique, construit à partir de fichiers annotés au préalable et de l'outil Stanford NER (appelable dans GATE).

Description du projet

Nous souhaitons enrichir cette chaîne de traitements dans GATE en ajoutant plusieurs fonctionnalités :

- désambiguïsation des noms propres de lieux. Par exemple, pour la chaîne *Paris* figurant dans le texte, les dictionnaires proposent plusieurs annotations correspondant à la ville *Paris* (France), la ville *Paris* (Texas), le mont *Paris* (Antarctique), etc. Différentes heuristiques existent dans la littérature pour choisir l'annotation correcte qui mixent des indicateurs spatiaux, linguistiques, statistiques.

- transformation d'annotations en TEI en annotations XML GATE. Pour certaines utilisations, le texte doit être annoté en format TEI mais la chaîne de traitements GATE pose des annotations XML. Il est nécessaire de passer d'un format à l'autre, dans les deux sens ;

[1] https://gate.ac.uk/

[2] Cette chaîne a été développée dans le cadre du projet et du projet "Des récits à la carte" (projet canadien avec lequel nous collaborons à travers une mission d'expertise).

- développement de métriques mesurant l'écart entre deux chaînes de caractères dans GATE. Il existe de nombreuses distances (Levenshtein, Hamming, Jaro, Jaro-Winckler, etc.) mais elles ne tiennent pas compte de phénomènes linguistiques observés et décrits dans la littérature tels que la troncature qui permet de passer de *Fontenay-aux-Roses* (nom officiel figurant dans le dictionnaire des lieux noms propres) à *Fontenay* (formulation trouvée dans des textes trouvés sur le Web). Il faudrait implémenter ces distances dans GATE, les tester dans différents corpus de textes, et proposer des modifications pour tenir compte de certains phénomènes linguistiques.

- calcul de l'emprise géographique d'un texte. Certains textes (en particulier les articles de journaux) contiennent de nombreux noms de lieu mais tous ces lieux ne sont pas importants par rapport au message véhiculé par le texte. Il s'agit de définir, grâce à des indices spatiaux, linguistiques, statistiques calculés à partir du texte, l'emprise géographique qui contienne les lieux importants du texte.

Ajouter une fonctionnalité correspond à du développement en JAVA puis une intégration dans GATE. Certaines de ces fonctionnalités ont déjà été développées en JAVA ; il faudrait les tester, les améliorer et ensuite les intégrer à GATE. D'autres fonctionnalités sont à développer *ex nihilo*.

Encadrement

IGN/Laboratoire LaSTIG : Catherine Dominguès, chargée de recherche
catherine.domingues@ign.fr

ESIEE : à préciser