Une chaîne d'analyse des e-mails pour l'aide à la gestion de sa messagerie

Gaëlle Recourcé[1]
(1) Kwaga SAS, 15 rue J-B. Berlier, 75013 Paris, France
recource@kwaga.com

Au sein de la société Kwaga, une équipe d'ingénieurs a réalisé une chaîne d'analyse des e-mails utilisée dans des applications d'aide à la gestion de sa messagerie pour les particuliers et les professionnels. Ce cœur technologique intègre au travers d'un chaînage UIMA, plusieurs composants de TAL issus de la recherche universitaire ou réalisés en interne.

1 Une chaîne d'analyse des mails

1.1 Import des messages

La première étape de la chaîne d'analyse des mails consiste à se connecter à un serveur IMAP pour importer une boîte mail (messages entrants et sortants). Les mails sont caractérisés par une en-tête (expéditeur(s), destinataire(s) directs ou en copie, date), un contenu (sujet et corps) et par leur organisation en conversations : un ensemble de messages échangés sous un même sujet par un ensemble de participants peut être considéré comme une séquence de répliques dans une conversation intégrant des apartés (transferts).

1.2 UIMA – annotation des mails

Le corps de la chaîne d'analyse de Kwaga est implémenté dans le cadre d'une séquence d'annotations UIMA (*Unstructured Information Management Architecture)*. Cette chaîne se subdivise en trois étapes, la détection du corps textuel, l'analyse linguistique, et l'interprétation.

1. Par le jeu des réponses et transferts, le corps d'un e-mail se structure par des niveaux de reprises successifs, indiquant le degré de nouveauté dans la conversation. Cette première annotation (CAS) consiste à repérer le texte nouveau, i.e. effectivement produit par le dernier expéditeur dans la conversation. Ce texte est par ailleurs soumis à la reconnaissance de langue (grâce à un module adapté de TextCat) et diverses expressions régulières sont appliquées sur les champs structurés et sur le corps de message permettant de calculer des informations caractérisant le message qui seront utilisées dans la phase d'interprétation.

2. La phase d'analyse linguistique est réalisée par l'application de graphes sur le corps et le sujet du mail par le biais d'une librairie JNI d'Unitex. Ces grammaires locales permettent d'une part de repérer les éléments de la structure du message (formules introductives, salutations finales, et signatures) et d'autre part des éléments du texte utilisés pour l'interprétation : phrases prototypiques (demandes d'action, proposition de rencontre, facturation…) ou éléments d'information assimilables à des entités (dates, mots de passe...). Ces automates sont appliqués en une passe unique, les sorties constituant une nouvelle annotation du CAS UIMA.

3. Dernière phase d'annotation UIMA, l'interprétation exploite les informations contextuelles (data, expéditeurs, destinataires, …) et les combine avec les indices linguistiques découverts dans le corps du texte, la signature ou le sujet pour calculer la catégorie du mail et les éventuelles informations associées telles les informations de contact. Ces informations dépendent aussi du contexte de conversation : un mail en réponse peut, dans certaines conditions, hériter certaines propriétés du mail qui l'a précédé.

2 Démonstrations (en ligne)

- Analyse d'e-mails à la demande (en français ou en anglais).
- Présentation de la création à la volée de corpus d'e-mails (serveur IMAP Gmail)
- Extraction d'information dans les e-mails – factures, mots de passe, fiches contact
- Catégorisation des e-mails : mail importants et Bac'n.