Axe 5
Structuration et Exploitation de Corpus (SEC)

Responsables d'axe : Mohamed Morchid MCF/UAPV et Graham Ranger PR/UAPV
Chercheurs impliqués (nom, prénom, discipline) : Bassam Jabaian (Informatique), Georges Linarès (Informatique), Raphaël Roth (Sciences de l'Information et de la Communication), Rachid Elazouzi (Informatique)
Mots clés : Structuration; Modélisation; Exploitation; Apprentissage automatique

1. Contexte

La structuration de l'information contenue dans un document donné consiste à extraire un ensemble fini de descripteurs de plus ou moins haut niveau ou des représentations concises et caractéristiques de ce contenu. Cette structuration permet de traiter de grands ensembles de données souvent connues sous le nom de "Big Data", depuis un ensemble réduit d'éléments caractéristiques des contenus. Cette tâche de structuration se situe en amont de processus de traitement de l'information de plus haut niveau tels que la catégorisation, l'extraction d'information ou l'indexation.

Il est donc nécessaire de définir des procédures de traitement de l'information permettant d'homogénéiser la gestion (format des données, processus de sauvegarde et de diffusion, définition d'éléments caractéristiques, etc.) ainsi que la diffusion de bases de connaissances provenant de disciplines scientifiques hétérogènes (histoires, littérature, culture, géographie…)

2. Objectifs

L'objectif premier de l'axe 5 SEC est de penser des modèles et méthodologies innovants de structuration et d'exploitation de l'information. Ces méthodologies seront fondées sur des paradigmes théoriques de modélisation et d'extraction de descripteurs de haut-niveaux pour le volet "structuration", ainsi que sur des procédés et processus de recherche d'information (RI) pour la partie "extraction" et valorisation des contenus. Ces nouvelles méthodologies seront évaluées lors d'expérimentations de traitement de l'information portant sur des corpus de données interdisciplinaires ainsi que sur des tâches dédiées..

Les chercheurs d'Agorantic travaillent sur des données dont les quantités et les structures internes sont très variables et spécifiques aux disciplines dont elles sont issues. L'axe 5 se propose alors de travailler sur l'accompagnement des membres d'Agorantic dans la préparation de données en vue de leur interrogation, leur modélisation et leur exploitation, via une interface mutualisée et mutualisable en ligne.

L'axe SEC mettra également en place des structures et des procédures qui permettront de promouvoir et de faciliter le travail interdisciplinaire au sein d'Agorantic, et aura comme objectif de promouvoir les inter-échanges entre les différents chercheurs de la structure en

organisant des rencontres (séminaires, workshops, etc.) autour de cette gestion commune de l'information.

3. Questionnements

L'axe 5 SEC aura pour vocation de répondre aux interrogations liées à la pluridisciplinarité dans l'espace digital :

→ Quelle est la place des grandes bases (Big Data) de connaissances dans les échanges interdisciplinaires ?

→ Comment définir un environnement de travail unifié permettant de traiter des problèmes connexes portés par des disciplines hétérogènes ?

→ Quelle capacité ont les paradigmes et méthodologies liés à l'apprentissage automatique à coder et extraire de l'information pertinente dans un contexte de pluri-disciplinarité ?

→ Les contraintes avec le CNIL. Une charte régissant les conditions d'accés et d'utilisation de la base de connaissance via la plateforme (site web) de mutualisation et de gestion des données sera disponible sur le site web.

4. Objets d'étude

Les modèles et méthodologies seront étroitement liés au domaine de l'apprentissage automatique. Parmi celles-ci, les espace thématiques permettent d'extraire des descripteurs abstraits de haut-niveau caractéristiques du contenu du document (thèmes, relation latentes entre les mots, mots-clés, etc.) comme indiqué dans la figure ci-dessous.
L'axe SEC exploitera également pleinement le potentiel très prometteur des représentations et systèmes fondés sur les réseaux de neurones profonds dans un premier temps, puis proposera des modèles théoriques et méthodologies innovants adaptés aux problématiques liées à la SFR Agorantic portées par le vecteur de la pluridisciplinarité. Parmi ces méthodes fondées sur les réseaux de neurones, les réseaux de neurones récurrents (RNN) permettent l'extraction de descripteurs robustes considérant la structure temporelle du flux d'information [1,6]. Les chercheurs du LIA travaillent sur ces problématiques de structuration de l'information pour le traitement automatique du langage [2,3,4,5]. Ces algorithmes permettent de traiter aussi bien de grands corpus de données (BigData) via des réseaux de neurones parcimonieux [5] que des données multidimensionnelles (texte+parole, texte web+texte littéraire, etc.) [2,3,4].

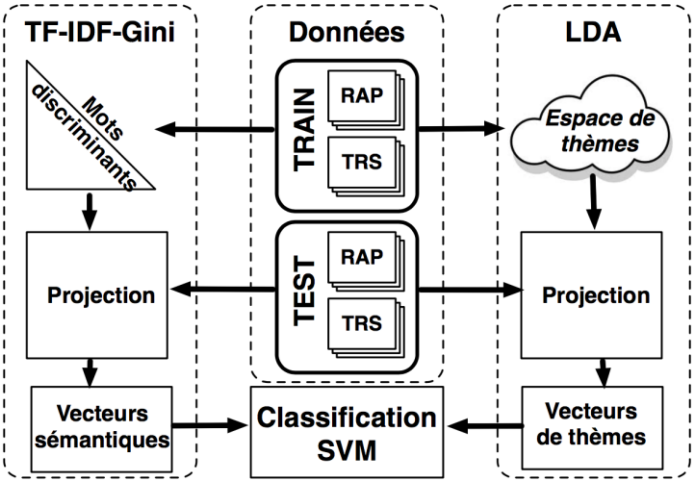

Figure : Exemple de projection d'un document dans un espace de thèmes (Latent Dirichlet Allocation ou LDA) pour l'extraction de descripteurs de haut-niveaux pour la classification de textes.

5. Nature des données à exploiter

Les données disponibles seront de nature très hétérogène et répondront à un besoin à la fois collectif (inter-disciplinaire) et dédié aux champs de recherche (à la discipline). Les données prennent des formes variées selon le médium et le média employé :

- textes issus du signal de parole
- documents textuels issus de plateforme d'échanges ou de microblogging (tweets, wikipedia, etc.)
- images
- etc.

Exemple de document parlé (DECODA 20101206_RATP_SCD_0254) entre un utilisateur de la RATP et un agent. L'agent à pour objectif de rediriger l'appel vers le bon sevice, ici "OBJET PERDU". Cet exemple illustre la difficulté d'extraire des descripteurs robustes pertinents depuis des transcriptions de documents parlés, liée aux nombreuses erreurs de transcriptions (système de reconnaissance de la parole, environnement bruité, téléphone portable, etc.)

Transcription issue du signal de parole

va vous répondre va vous répondre bonjour moi il me semble monsieur c'est en fait c'est je sais mais c'est à eux est-ce que vers treize heures j'ai pris le métro à un arrêt mais c'est oui j'ai les perdu mon portefeuille avec eux mais d'accès alors en aura des papiers à quel nom

Transcription humaine

va vous répondre bonjour oui bonjour monsieur en fait j appelle parce que euh c vers treize heures j ai pris le métro à à Arts-et-Métiers oui et j ai perdu mon portefeuille avec tous mes papiers dedans

alors les papiers à quel nom [...]

Références

[1] A.Graves,"Neural networks", in Supervised Sequence Labelling with Recurrent Neural Networks. Springer, 2012, pp. 15–35.

[2] M. Bouaziz, M. Morchid, R. Dufour, G. Linares, R. De Mori, "Parallel Long Short-Term Memory for multi-stream classification", IEEE Spoken Language Technology Workshop 2016.

[3] T. Parcollet, M. Morchid, P.-M. Bousquet, R. Dufour, G. Linares, R. De Mori, ”Quaternion Neural Networks for Spoken Language Understanding”, IEEE Spoken Language Technology Workshop 2016.
[4] T. Parcollet, M. Morchid, G. Linares, ”Quaternion Denoising Encoder-Decoder for Theme Identification of Telephone Conversations”, ISCA Interspeech 2017.
[5] M. Morchid, ”Internal Memory Gate for Recurrent Neural Networks with Application to Spoken Language Understanding”, ISCA Interspeech 2017.
[6] K. Janoid, M. Morchid, R. Dufour, G. Linarès, R. De Mori ”Denoised Bottleneck Features from Deep Autoencoders for Telephone Conversation Analysis”, IEEE Transaction on Audio, Speech, and Language Processing, 2017.