# Axe Accès à l'information et fouille de textes

Thierrry Charnois et Jean-Pierre Chevallet

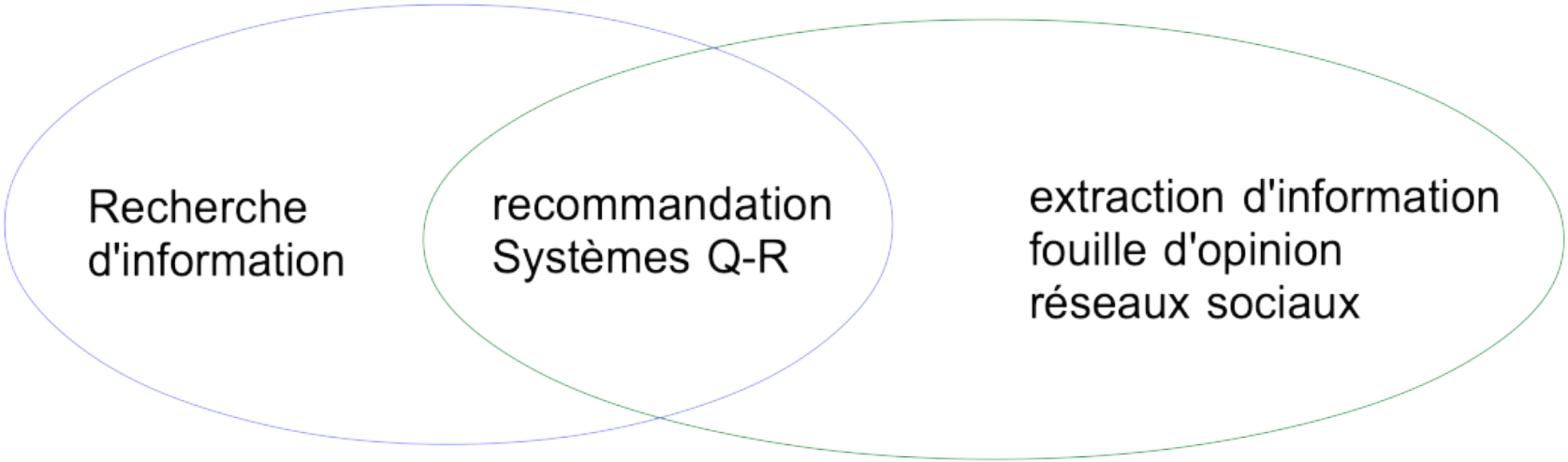
Pré-GDR TAL

24 octobre 2019

# Participants

LIS Marseille
LIMSI Orsay
LIG Grenoble
IRIT Toulouse
LIPN Paris-Nord
IRISA Rennes
LORIA Nancy
L2SN Nantes
GREYC Caen
CEA List

# Périmètre




**Méthodes**
- apprentissage,
- fouille de graphes
- ...

**Domaines**
- BioNLP
- TAL juridique
- SHS
- réseaux sociaux
- retranscriptions oral
- .
- .
- .
- Open domain

**Connaissances**
**Ressources**

# Extraction d'information : état des lieux

- But : extraction d'éléments "pertinents" dans les textes (REN, relations entre EN, événements, relations entre événements...)
- Utilisé pour l'enrichissement de base de connaissances ou comme préalable à d'autres tâches de TAL
- 2 cas d'usage
  - extraction d'infos dont le type est prédéfini (repérage d'instances d'entités, de relations)
  - exploration : découverte d'entités et de relations (OpenIE)
- Méthodes : prédominance des méthodes d'apprentissage (réseaux de neurones, clustering)

# Extraction d'information : défis

- Précision des infos extraites : type d'entités plus fin (acteur, politicien vs personne), désambiguïsation (Entity Linking), extraction des relations et événements (F-score 56% à TAC KBP)

- Quantité des données annotées
  - apprentissage supervisé (réseaux neurones)
  - faiblement supervisé (supervision distante avec bases de connaissances)
  - non supervisé : problématique de l'Open IE
- Adaptation automatique des modèles d'extraction d'information au domaine ou au genre, pour prendre en compte l'hétérogénéité des textes (genres, registres de langue, oralité, réseaux sociaux avec les sociolectes...)

# Fouille d'opinion : état des lieux

- Problématique
  - de la classification supervisée de sentiments (initialement polarité) et plus récemment avec des RN
  - évolution vers des analyses de + en + fines : polarité mais aussi émotions, cibles des émotions
- de nombreuses applications : recommandation, veille sur la réputation d'entreprise....

# Fouille d'opinion : défis

- détection d'émotions spécifiques (peur, rage) + intensité, distinction émotion suscitée / émotion du rédacteur
- identification de langage figuratif (sarcasme, ironie)
- détection d'attitude ("stance") : position du locuteur par rapport à la situation énoncée
  – *orateur A : les migrants ont le droit à un meilleur endroit pour vivre.*
  – *orateur B : c'est pour ça que 25% de la population pénitentiaire est composée d'étrangers...*

  B est contre la position de A –> tâche difficile (inférence et connaissances nécess.)
- adaptation au domaine, genre (e.g. tweets) –> cf. défi IE

# Analyse des réseaux sociaux : état des lieux

- Analyse
  - contenus échangés
  - liens entre individus ou communautés
- Nombreuses applications : détection de changement chez un individu (e.g. agressivité, challenge TRAC), extraction de communautés sur le même sujet, études des influences sur un sujet...

# Analyse des réseaux sociaux : défis

Liés aux 2 types d'infos (contenus, liens)

- exploiter infos textuelles pour l'analyse des interactions –> utiliser représentation des textes dans algos d'analyse des graphes d'interaction
- réciproquement, exploiter les infos sociales pour l'analyse des textes –> ex. détection de commauté pour connaître le pt de vue d'une personne (ironie, relativité du sens selon ce pt de vue)
- prise en compte des spécificités de la langue (voc., types de communication, emojis, hastags...), hétérogénéité des données (nature des interactions peut varier selon nature du réseau)
  - –> méthodes robustes génériques
  - –> adaptation automatique des modèles à ces données

# Recommandation : état des lieux et défis

- But : proposer des items à l'utilisateur (avec ou sans requête)
- Méthodes :
  - partagées par RI (RI orientée recommandation)
  - analyse des activités (eg traces)
  - analyse des retours utilisateurs (note, évaluation textuelle)
- Défis (TAL, dialogue, RI, opinion)
  - développement assistants vocaux nomade et personnels (technos "voice control")
  - analyse sentiments et émotions
    - de l'utilisateur vis à vis d'un item
    - sur le contenu même (livre triste ou joyeux ?)
  - mise en correspondance caractéristiques souhaitées / trouvées (RI et pertinence multi-dimensionnelle), traitement requêtes longues en langue

# Première synthèse

- La fouille de textes un processus d'exploration des textes à grande échelle et hétérogènes
- Les défis de la fouille de textes :
  - prise en compte de l'hétérogénéité des textes
  - adaptation des modèles ou des méthodes d'apprentissage
  - analyse en domaine ouvert, découverte de connaissances : apprentissage non supervisé
  - affiner les analyses : question de l'intégration, l'acquisition, modélisation de ressources et connaissances –> lien fort avec l'ingénierie des connaissances
- l'exploitation des données langagières : un enjeu important pour la recherche et l'innovation :
  - analyse de tendances, cartographie d'un domaine, recherche d'experts, émergence de notions latentes, recommandation,
  - champ des Humanités Numériques

# Recherche d'information : état des lieux

TALN pour les SRI, vers des SRI 'sémantiques', i.e. exploitant des ressources.

- Quelle "sémantique" est utile et exploitée par les SRI ?
- Quelles sont les ressources textuelles qui sont utiles et effectivement utilisées dans les SRI (ex : ontologie, terminologie, thésaurus) ?
- Quel est l'utilisation effective des ressources dans les SRI ?
- Comment évaluer l'impact de l'usage de ces ressources sur la qualité finale du SRI en terme de satisfaction de l'utilisateur ?

# Recherche d'information : typer les ressources

- Ressources de domaine (en médical UMLS, Gene Ontology, etc),
- Ressources générales (Wikipedia, dbPedia, Garo, Geonames, Wordnet, etc),
- Ressources issues de service web (Foursquare, Telp, etc),
- Ressources pour la sémantique distributionnelle. Ex : Construction et exploitation de plongements de mots (word embedding)

Mais aussi : classification des ressources en fonction de leur niveau de formalisme (i.e. ontologie + formalisme logique)

# Recherche d'information : exploiter les ressources

- Les approches orientés appariement : améliorer la représentation des requêtes ou des documents pour augmenter les chances d'appariement (term mismatch) :
  - expansion de la requête : reformule la requête à l'aide de mots / concepts pertinents, ou on pondère les termes de la requête ;
  - expansion de document :expansion statique, i.e. indépendante de la requête, ou dynamique, en tenant compte de la requête ou du retour utilisateur ;
  - expansion de la requête et des documents.
- modifier l'ordonnancement ;
- apprentissage : d'apprendre la fonction de pertinence, ou apprendre un représentation des mots / concepts / documents. => Deux sous domaines : le 'learning to rank' et les approches de correspondance par des réseaux de neurones.

# Recherche d'information : verrous

- Qualité des ressources vs efficacité en RI sémantique ?
- Niveaux de sémantique plus élevé : sémantique du dialogue (systèmes conversationnels), sémantique de la tâche/session (systèmes orientés tâches/décision, systèmes de RI aggrégative générant des objets à valeur ajoutée, etc.)
- Sémantique transparente et explicable : Au delà de l'utilisation de représentations textuelles "sémantisées", qualité des "inférences" sémantiques les expliquer à l'utilisateur (Fairness and transparency in IR)
- Sémantique dans différents domaines de l'informatique : RI, Machine Learning, Gestion de Connaissances, NLP : quelle sémantique ? aboutit-on à des 'sens' alignables ? Quelles différences et pourquoi ?
- Nouvelles formes de RI : dialogue.

# Recherche d'information : modes, tendances ... et IA

Exploitation de techniques d'apprentissage automatique :

- Apprentissage non supervisé : construction de plongements de mots (word2Vec, BERT, ...), RI à l'aide de ces structures apprises.
- Modèles de RI à base de réseaux de neurones profonds, mais après le filtrage par une SRI classique => ré-ordonnacement
- Modèle de RI "end to end" : apprendre aussi des structures creuses.
- Inclusion de ressources (discrète) dans des structures continues et leur exploitation "end to end".

Questions :

- Des corpus RI assez larges pour l'apprentissage "end to end" ?
- Est-ce que l'apprentissage sur des ressources langagières apporte quelque chose de décisif à la RI ?