Quel TAL pour le retour d'expérience ?
Réflexions sur les besoins et les solutions actuels

Ludovic Tanguy

CLLE-ERSS : CNRS & Université de Toulouse

IMdR – 16 Mai 2017

PLAN

- Les REX vus du TAL
  - Caractéristiques des données et du contexte
  - Vue d'ensemble des applications
- Le TAL et ses évolutions
  - Modes de représentation d'un texte
  - Des ressources a priori aux techniques inductives
- Perspectives
  - Les nouvelles techniques de TAL pour les REX
  - Contraintes de la recherche en TAL

LES REX VUS DU TAL : SPÉCIFICITÉS, BESOINS, APPLICATIONS

- Caractéristiques des REX (1)
  - Volume
    - *masse critique atteinte*
      - 10-100 K de docs, 1-10 M de mots pour les approches statistiques
    - *manipulable sans infrastructure démesurée*
      - Les collections tiennent sur un disque dur standard
    - *« Bonne taille » pour le TAL actuel*
  - Hétérogénéité et qualité de l'expression
    - *Variation en taille et types de documents, texte saisi à la volée et grand nombre de scripteurs*
    - *TAL aguerri depuis longtemps grâce aux données issues du Web et des réseaux sociaux*
  - Métadonnées
    - *Riches, souvent systématiques et parfois fiables*
    - *Permettent le croisement avec le contenu textuel et le traitement différencié*

- Caractéristiques des REX (2)
  - Langue de spécialité
    - *Difficulté à projeter des modèles/ressources génériques*
    - *Nécessité ou bénéfice de ressources terminologiques spécifiques (ou de leur construction)*
    - *Problématiques classiques depuis le début du TAL*
  - Normes et référentiels
    - *Organisation des métadonnées*
    - *Modèles en amont et/ou en aval*
    - *Normalisation des contenus*
      - A intégrer dans les traitements
      - A contrôler dans certains cas
    - *Liens entre le TAL et l'ingénierie des connaissances*

- Caractéristiques des REX (3)
  - Rapport incontournable à l'expert
    - *Rôle central dans les processus autour des REX*
    - *Principal destinataire des traitements*
    - *Attribution des métadonnées interprétatives*
    - *Construction/validation des ressources*
    - *Eclairage utile à toute étape d'une chaîne de traitement*
    - *Besoin de penser son intégration à une application de TAL*

- Quels besoins (Blatter & Raynal, projet P10-5) ?

Fouille de données textuelles

tous

Catégorisation automatique

Roussel & Latchurie ; Gaucher, Chami & Vasseur ; Marle et al. ; Quéva ; Martin & Reutenauer

Recherche d'information

Marle et al. ; Martin & Reutenauer ; Blatter & Quéva

Calcul de similarité entre documents

Marle et al ; Quéva ; Simon-Biscarat et al.

Clustering

Tissot

Constitution de ressources langagières

Roussel & Latchurie ; Marle et al. ; Blatter & Quéva

Extraction d'infos sémantiques spécifiques

Blatter & Quéva

Modélisation fine du contenu

Blatter & Quéva

- Au final, des besoins classiques en TAL
  - Recherche d'Information, catégorisation, clustering, Extraction d'Information, fouille de textes
- Question centrale : la représentation du texte
  - Modélisation
    - *Calcul de similarité entre documents*
    - *Recherche d'information*
    - *Classification*
  - Traits génériques d'un texte
    - *Intégration dans le data-mining*
  - Structures locales
    - *Terminologie*
    - *Extraction d'information*

LES ÉVOLUTIONS DU TAL

- Tournant des années 2000
- Avant :
  - Culture de l'IA symbolique et de la linguistique formelle : modélisation précise du contenu
  - Analyses fines et locales
  - Ressources linguistiques et connaissances externes
- Après :
  - Modélisation statistique du langage
  - Apprentissage automatique
  - Acquisition d'information de moins en moins supervisée

Exemple de modélisation

Le passage du diesel par la voie de sortie du dépot n'étant pas possible, l'AC du poste K décide de faire sortir l'engin moteur par mouvement à contre-voie en gare par l'itinéraire d'entrée.

```
[Mouvement: y]
        -->(FP)-->[Fonction Principale: #]
                        -->(supportée par)
                                -->[Véhicule: diesel]
                                -->[Itinéraire:Voie sortie]
        -->(type)-->[#]
        -->(destination)-->[#]
        -->(responsable)-->[#]

(impossible)-->[Mouvement: x]

[Décide]-->(agent)-->[Agent de circulation]-->(identificateur)-->[Poste K]
         -->(objet)-->[Mouvement: y]

[Mouvement: y]
        -->(FP)-->[Fonction Principale: #]
                        -->(supportée par)
                                -->[Véhicule:Engin moteur]
                                -->[Itinéraire:Voie entrée]
        -->(type)-->[contre voie]
        -->(destination)-->[gare]
        -->(responsable)-->[Agent de circulation]-->(identificateur)-->[Poste K]
```

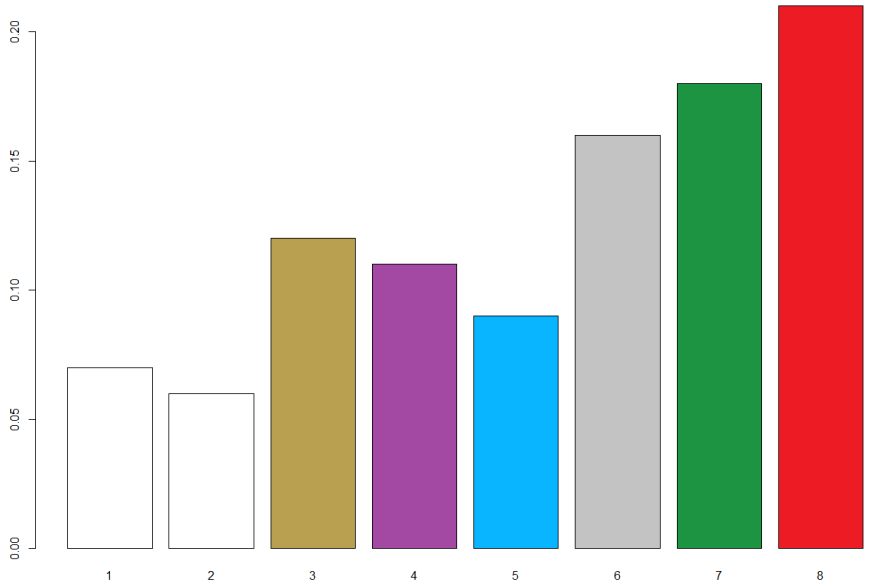
Lannoy et al. (1992)
Analyse automatique de textes libres – projet 12/91. *Institut de sûreté de fonctionnement.*

- Caractéristiques des approches « fines »
  - Objectifs ambitieux
    - *Modélisation fine du contenu*
    - *Forme de raisonnement*
  - Ressources explicites et complexes
    - *Analyse linguistique locale complète*
    - *Lexiques et terminologies spécialisés*
    - *Connaissances formalisées du métier et du domaine*
  - Objectifs rarement atteints
    - *Développement et évolution des ressources*
    - *Robustesse face aux données non-normées, incomplètes, en évolution*

(Emprunté à Tissot)

Un ouvrier de production de 25 ans, employé en contrat à durée déterminée, préparateur des charges sur la machine ***, est en cours de formation pour devenir conducteur polyvalent. Son travail consiste principalement à surveiller l'alimentation et à accompagner la bande de matière sur le tablier existant entre le mélangeur et le metteur en plaques, pour pallier au manque d'adhérence. A l'examen des circonstances de l'accident, il semble que l'ouvrier muni de gants accompagnait la gomme de la main gauche pour favoriser son introduction dans le metteur en plaques. Au cours de cette opération, l'extrémité de son gant a été happée par l'angle rentrant constitué par une courroie et le cylindre d'entraînement de la gomme. Alerté par les cris de l'opérateur dont l'avant-bras gauche était engagé dans le metteur en plaques, un collègue a immédiatement actionné le dispositif coup de poing d'arrêt d'urgence. Souffrant d'une fracture ouverte de l'humérus, l'ouvrier a été hospitalisé.

Thème 1 : chute de hauteur
Thème 2 : manutention, levage
Thème 3 : réglementation
Thème 4 : circulation, chargement
Thème 5 : procédures
Thème 6 : risque électrique
Thème 7 : accueil, formation
Thème 8 : risque machine



- Caractéristiques des approches « stats »:
  - Objectifs réduits
    - *Catégorisation, identification des thématiques globales d'un document*
  - Modélisation plus grossière du langage
    - *Mots ou termes en cooccurrence*
    - *Approche fréquentielle*
    - *« Données non structurées », « sacs de mots », etc.*
  - Robustesse face aux données
  - Besoins réduits en ressources et connaissances
    - *Mais exigence d'un volume minimal de texte*

- Changement de la représentation d'un texte
  - D'un objet complexe, structuré, composé d'unités diverses et liées entre elles
  - A un vecteur
- Représentation vectorielle
  - Document représenté par un ensemble de coordonnées numériques
    - *Basées essentiellement sur le contenu lexical*
  - Applications
    - *Mesures de proximité classiques type cosinus*
    - *Compatibilité avec les métadonnées*
- Quel espace vectoriel ?

Mots	*camion*	*voiture*	*arrêter*	*bloquer*	*chaussée*	*accès*	*secours*
Camion arrêté sur la chaussée	1	0	1	0	1	0	0
Voiture bloque l'accès aux secours	0	1	0	1	0	1	1

Classes de mots	*camion, voiture, moto, véhicule…*	*arrêter, bloquer, stopper…*	*chaussée, rue, route…*	*accès, entrée, porte…*	*secours, assistance, aide…*
Camion arrêté sur la chaussée	1	1	1	0	0
Voiture bloque l'accès aux secours	1	1	0	1	1

Dimensions abstraites	D1	D2	D3
Camion arrêté sur la chaussée	0,78	0,43	0,9
Voiture bloque l'accès aux secours	0,82	0,24	0,1

- Dimensions induites
  - Principe : repérage des comportements similaires des unités
  - Techniques phares :
    - *Sémantique latente (LSA, LSI) : Deerweister et al. 1990*
    - *Topic modeling (LDA) : Blei et al. 2003*
    - *Sémantique distributionnelle (word embeddings) : Mikolov et al. 2013*
  - Résultats :
    - *Dimensions en nombre réduit*
    - *Calcul non supervisé, basé sur corpus*
    - *Définition plus robuste de la similarité*
    - *Dimensions non ou difficilement interprétables*
      - (sauf examen des termes ou documents rapprochés)
      - Et donc similarité ou classification non « explicable »

- *Topic modeling*
  - Regroupement parallèle documents / termes
  - Exemples de classes induites sur corpus ASRS (outil Mallet) :
    - *Lights, light, panel, night, lighting, illuminated, red, green, bright, dark, vision, overhead, eye, sun, flashing…*
    - *Frequency, radio, tower, contact, communication, hear, atc, frequencies, communications, heard, response, radios…*
    - *Day, hours, time, rest, duty, night, crew, fatigue, hour, sleep, trip, flight, morning, scheduling, call, scheduled, late, zzz…*
  - Limites :
    - Principales thématiques des documents et de la collection
    - Difficulté d'interprétation des classes
    - Grande variabilité des résultats suivant les paramètres (nombre de dimensions)

- Analyse distributionnelle
  - *Word embeddings (plongements lexicaux)*
  - Représentation vectorielle réduite des unités lexicales
    - Sur la base de leur contexte local
  - Permet le calcul d'une similarité lexicale
    - Exemple sur corpus ASRS (Word2vec) :
    - *fatigue : complacency, distraction, scheduling, human, alertness…*
    - *smoke : smell, fumes, odor, flames, cabin, vents, noxious*
    - Capte des relations sémantiques diverses (synonymie, antonymie, hyperonymie, co-hyponymie) mais imprécises
  - Extension aux unités supérieures (phrases, documents)

- Utilisation directe
  - Aide à la construction de ressources terminologiques
- Utilisation dans les applications comme représentation du lexique
  - Systèmes de classification / clustering / indexation
  - En entrée des outils de traitement des données
    - *Traduction, étiquetage, parsing, etc.*
  - Bénéfices
    - *Gestion économique de la variation lexicale*
    - *Données compactes et calcul plus rapide*
    - *Gain d'efficacité mesuré selon les standards actuels*
    - *Approche non-discrète de la sémantique*

- Conséquences sur le courant actuel (Halevy et al. 2009)
  - Elimination des ressources en amont
    - *Juste de grands volumes de textes (plus ou moins sélectionnés)*
  - Limitation des interventions dans les prétraitements
    - *Plus de « feature engineering »*
  - Justifications :
    - *Coût réduit, absence de biais, gain en fiabilité et en couverture*
- Philosophie du *deep learning* (Manning 2015)
  - Architectures neuronales complexes
  - Phase 1 : apprentissage non supervisé sur de gros volumes de textes non traités et non labellisés
    - *Word embeddings et autres*
  - Phase 2 : apprentissage supervisé sur un petit volume

RETOUR AU REX

- Quelle place pour ces nouvelles techniques ?
- Intégration déjà bien engagée
  - Telle quelle dans les approches vectorielles
  - Induction de classes de documents/termes
  - Idéal pour les applications sans référentiel ni ressources a priori
    - *Pour les construire ou s'en passer*
- Questions ouvertes :
  - Quelle interaction avec les connaissances/modèles/ressources existants ?
    - *Sur la langue et sur le domaine*
  - Quelle pertinence pour les phénomènes rares ?

- Quelle place pour l'expert dans le développement des systèmes ?
  - (et le linguiste)
- Approches classiques
  - Participe à la construction des ressources
    - *Terminologies, ontologie, structures locales*
  - Valide les résultats
    - *Identifie et diagnostique les erreurs*
  - Participe aux traitements intermédiaires
    - *Oriente ou définit les règles, affine les réglages*
    - *Interprète les configurations complexes ou atypiques*
- Approches par apprentissage
  - Labellise ou classe les données
    - *En entrée pour construire le modèle*
    - *En sortie pour mesurer l'efficacité*

- Limite à l'interaction avec l'expert :
  - Opacité du calcul
    - *Impossibilité de retracer les calculs à partir des résultats*
    - *Difficulté à identifier les pistes d'amélioration/correction*
  - Manque de moyens d'action sur le système
    - *Seul levier : l'injection de nouveaux cas d'apprentissage*
    - *Cf. apprentissage actif (PLUS)*
  - Aveuglement aux phénomènes rares ou atypiques
    - *Signaux trop faibles*
  - Coût réduit
    - *Difficile à ignorer si la performance est comparable*
    - *Difficile d'évaluer le bénéfice d'un expert qui a une expérience directe avec les traitements des données*

- Pistes de solution
  - Intégrer des connaissances en amont
    - *La projection de ressources reste possible malgré tout*
    - *Terminologies, choix des contextes, structure du document…*
  - Multiplier les techniques
    - *Pas de méthode unique face aux questions*
    - *Exemple : dégrossir avec des traitements robustes, identifier et traiter les cas difficiles par des méthodes locales*
  - Former à la complexité accrue des outils
    - *Triple culture des utilisateurs : domaine, TAL, machinerie statistique*
  - Recourir à la visualisation
    - *Et à l'interactivité*

- Quel rôle pour la recherche académique ?
  - *Les REX et la gestion du risque sont des domaines motivants et exemplaires (et aussi utiles)*
  - *Mais pas toujours bien connus ni accessibles*
  - Quelques obstacles
    - *Accès aux experts*
      - Intensif, surtout si on expérimente
    - *Confidentialité des données*
      - Reproductibilité et ouverture des données de plus en plus fréquemment exigées dans les publications
    - *Evaluation*
      - Les jeux de test classiques du TAL ne sont pas pertinents quand on travaille sur des données spécialisées
      - Certains objectifs ne sont pas évaluables : identification de nouvelles thématiques, de corrélations pointues ou de signaux faibles
  - Pistes
    - *Projets collaboratifs avec les laboratoires universitaires*
    - *Tâches partagées*

Fabre, C., & Lenci, A. (2015). Distributional Semantics Today: Introduction to the special issue. *Traitement Automatique des Langues*, *56*(2), 7-20.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, *24*(2), 8-12.

Lannoy A. et al. (1992) *Analyse automatique de textes libres – projet 12/91*. Institut de sûreté de fonctionnement.

Manning, C. (2015). Computational Linguistics and Deep Learning. *Computational Linguistics* 41(4): 701-707

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* (3111-3119).

Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., & Raynal, C. (2016). Natural Language Processing for aviation safety reports: from classification to interactive analysis. *Computers in Industry*, *78*, 80-95.