RTR DIAMS - REUNION Groupe de Travail RITUEL sur le Traitement Automatique des Langues
Evaluation des technologies langagières - Evaluation du TAL en contexte industriel : quelles pratiques, quels besoins ? - Evaluation en contexte académique : quelles pratiques, quelles limites ?

LIEU Orléans, Hôtel Dupanloup, 1 rue Dupanloup 45000 Orléans
DATE 06 février 2020,10:45 - 17:30

COMPTE –RENDU

Compte-rendu rédigé par Jean-Yves Antoine avant tout à partir des notes précises et exhaustives de Claire Cailleau et Ilaine Wang.

Participants

AKTAN	Eric Clairambault
LISMI	Sharlene Lefevre
LIFAT	Jean-Yves Antoine, Adam Lion-Bouton, Ilaine Wang, Denis Maurel
LIFO	Andreanne Roques
LLL	Emmanuel Schang, Lotfi Abouda, Hélène Flamein, Flora Badin
BRGM	Claire Cailleau, Vincent Godard, Cécile Graciane
ATOS	Olivier Gracianne
QUAL'NET	Martial Relier
EDF	Julien Kahn
PRISME	Philippe Ravier

Présentation de la journée : quelle évaluation pour le TAL, entre évaluation techno-centrée et évaluation centrée utilisateur... Jean-Yves ANTOINE (LIFAT)

Le TAL et les entreprises :

- Sollicitations de plus en plus nombreuses de la part des entreprises sur des problématiques relevant du TAL ;
- Prise de conscience par les entreprises de la valeur présente dans les sources de données textuelles ;
- Souvent les problématiques exposées par les entreprises ne relèvent pas de problématiques de recherche mais de la mise en œuvre de techniques connues depuis une quarantaine d'années.

Introduction aux méthodes d'évaluation académique :

- Recherche de bonnes pratiques à partir des années 90 (avant pas ou peu d'évaluation)
- Evaluation avec méthodes inspirées des sciences physiques (métriques objectives proches de celles mises en œuvre dans le domaine du traitement)
- Benchmark des différentes méthodes sur des jeux de données identiques : évaluation de la performance du système à réaliser une tâche donnée => attribution de scores de performance

Exemple : reconnaissance des entités nommées. Tâche : délimitation des entités nommées + attribution d'un type (lieu, personne, produit..). Mesure : Slot Error Rate

- Les méthodes d'évaluation mises en œuvre génèrent des interrogations sur :
  - La représentativité et la fiabilité des jeux de données de référence
  - L'interprétation des résultats
  - La capacité des résultats à fournir des pistes d'améliorations des systèmes

Quelle place pour l'utilisateur dans l'évaluation des systèmes ?

- Existe-t-il une corrélation entre évaluation objective et jugement humain ?

  L'amélioration du score d'un système se traduit-elle systématiquement par une satisfaction utilisateur augmentée ? Pas nécessairement, par exemple en matière de traduction automatique / génération automatique de textes on observe une faible corrélation

- Comment expliquer ce constat contre-intuitif ?
  - Par l'artificialité de la tâche ? des données ?
  - En matière de prédiction de mots pour l'aide à la communication, on observe un meilleur score du système sous dictée/tâche prescrite qu'en conditions réelles /tâche naturelle (par ex. conversation)

Quelle évaluation ?

- L'évaluation en TAL est trop techno-centrée : elle permet bien la comparaison des systèmes mais est peu interprétable et contribue peu à l'amélioration des systèmes car elle est trop basée sur des tâches artificielles
  - Point fort : objectivation
  - Point faible : tâches et données artificialisées
- L'évaluation en ergonomie est plus anthropo-centrée
- La solution pourrait être de viser une voie moyenne en passant d'une logique benchmark à une logique test unitaire : privilégier un ensemble de jeux de tests se focalisant sur un phénomène, une difficulté précise

Exemple d'approche centrée utilisateur : contexte et besoins d'évaluation TAL pour les Règles Générales d'Exploitation Julien KAHN (EDF R&D)

Développement d'un outil d'aide à la recherche d'information dans le RGE (référentiel de règles d'exploitation) de l'EPR de Flamanville (intègre résultats du stage d'Aurore Hamimi)

RGE = code de la route des centrales nucléaires : document de référence

- 6000 p., 11 chapitres
- Structuré pour faciliter l'accès à l'information
- Traite des problématiques de sûreté, d'environnement, …
- Destiné à des utilisateurs distincts pour des besoins distincts
- Se caractérise par une langue technique + règles d'écriture sur les objets techniques
- Malgré cette normalisation : persistance d'ambiguïté en conditions réelles d'utilisation

- Utilisateurs cibles : équipes de conduite et ingénieurs sécurité

- Contexte d'utilisation : contrainte de temps
- Besoins :
    - Aider les utilisateurs à accéder à l'information : l'utilisateur, qui est un expert, doit pouvoir trouver des informations pertinentes et suffisantes pour faire la meilleure analyse possible et confronter cette analyse avec un tiers afin de prendre une décision… Il n'y a donc pas une seule et unique bonne réponse : les éléments à retrouver sont des éléments de réponse qui doivent aider l'ingénieur à prendre une décision
    - En cas de défaillance, permettre d'identifier de façon rapide et sûre les règles à mobiliser
    - Identifier des éventuelles incohérences dans les règles énoncées

Méthodologie : formulation de « scénarios types » d'usage du RGE

- Observation des modalités de recherche des utilisateurs
- Définition des exigences pour un système d'aide à la recherche d'information
    - Si génération d'une réponse automatique : nécessité forte d'explicabilité de la réponse proposée
    - Garantie d'exhaustivité de l'information utile contenue dans le RGE : focalisation sur le rappel et non la précision (de fait, le rappel doit être de 100%, on veut ne rien manquer).

Développement d'un démonstrateur technique (besoin rapide de démonstration)

- « Open semantic search » https://www.opensemanticsearch.org/ (+serveur Apache Solr) : système de recherché d'information symbolique à base de règles. Choix d'une approche par règles justifié par un besoin de démontrer le plus simplement possible les méthodes employées
- Etapes : Formatage > découpage > prétraitement > extraction d'information > recherche d'information
    - Découpage en « unités documentaires » signifiantes pour les utilisateurs (> 484 unités) – script Java
    - Prétraitement : Détection de « Repères fonctionnels » + normalisation des variantes de ces repères selon des règles explicables (car à justifier devant l'Autorité de Sûreté Nucléaire). Constat que certains repères échappent à la règle
        - Comment être sûr d'avoir identifié tous ceux qui échappent à la règle ?
        - Pour améliorer la captation des repères : mise en place de méthode de fouille avec détection de pattern et de méthode de « recherche floue » et de méthodes employées pour la correction orthographique
        - La normalisation ne doit pas nuire au sens

Critères d'évaluation du système

- Adéquation avec besoins utilisateurs
- Intelligibilité des réponses fournies
- Maintien des capacités des opérateurs à effectuer leurs recherches selon leurs modalités habituelles de recherche
- Exhaustivité : 100% nécessaire pour permettre l'utilisation du système
- Méthode projet : panel utilisateurs qui peuvent utiliser le démonstrateur en permanence, font remonter leurs retours à l'expert métier pour évolutions rapides

- Méthode d'évaluation technique de la performance du système :
    - Pour le moment, la méthode consiste à appliquer les règles définies sur un chapitre à un autre chapitre
    - Projet de comparer la détection obtenue par les règles avec la détection basée sur réseaux de neurones (grammaire locale vs fouille)
    - Besoin en évaluation : (1) adéquation aux besoins des utilisateurs (intelligibilité, compatibilité voire complémentarité avec compétence initial des utilisateurs à chercher dans le doc papier) avec comme contrainte de vraiment pouvoir garantir à 100% de pouvoir aider l'ingénieur sur certains éléments, et en plus le besoin de "circonscrire" les éléments garantis 100% (2) sur la normalisation de certains éléments, besoin de récupérer TOUTES les variante

Pratiques d'évaluation autour d'UKO Voice sur des exemples de projets clients Eric CLAIRAMBAULT (Aktan)

Aktan

- Cœur de métier : « service design » pour le développement de projets innovants, aide à la transformation des entreprises
- Approche UI/UX
- Clients grands groupes : trouver de nouveaux gisements de valeurs
- Campagnes d'observation sur le terrain des équipes => équipes d'ergonomes. Remontée d'observation sous formats divers : notes manuscrites, vidéos, audios, questionnaires
- Travaux pour outiller le traitement de ces données
    - Développement d'un modèle d'analyse multidimensionnel (Valeur Expérience Contexte Utilisateur)
    - Automatisation des opérations de traitement (dont extraction d'insights)

UKO Voice

Outil en cours de développement. « Voice » : terme employé pour marketer l'aspect « écoute de la voix de l'utilisateur »

Objectifs : mettre en évidence :

- Insights de conception : contraintes et habitudes
- Insights d'innovations et gisements de valeurs
    - Analyse de besoins au-delà de la stricte verbalisation

Développement :

- D'une interface pour prise en main par le client des données extraites des analyses
- D'un moteur d'analyse basé sur approche hybride croisant approche linguistique (règles et linguistique) et approche probabiliste (réseaux neurones)
- D'un « modèle d'apprentissage générique »
    - Modèles pré-entraînés (type Bert https://github.com/google-research/bert)
    - Travail sur le « transfert learning » / « fine tuning » -> mention du PD de S. Ruder (https://ruder.io/thesis/neural_transfer_learning_for_nlp.pdf). La question du transfert learning est importante, car les modèles généraux pré-entrainés sont peu adaptés aux besoins spécialisés d'Aktan. Le transfert learning ne suffit toutefois pas

nécessairement, d'où besoin de ré-entrainer un RNN ou un LSTM sur des données propres.

- D'un annotateur (format BIO/BILUO) utilisable en entrée par les réseaux de neurones (https://en.wikipedia.org/wiki/Inside%E2%80%93outside%E2%80%93beginning_(tagging) pour ce réentrainement. L'annotateur suit une approche centrée connaissance avec des automates locaux (Unitex : https://unitexgramlab.org/fr)

Voir aussi : https://www.quantmetry.com/bert-google-ai-banc-de-test/

Processus

- Extraction d'information :
  - Niveau sémantique
  - Détection d'entités, de concepts (contenu propositionnel), analyse de sentiment (polarité) dans une visée non pas d'analyse de sentiment aveugle et global, mais d' *Aspect based sentiment analysis* : on cherche à associer le sentiment à un « objet » : entité + entité-sentiment)
  - Ressources techniques employées
    - CasEn : détection d'entités
    - Unitext : création de grammaires et de grammaires spécialisées selon les métiers
    - Dependency parser : bibliothèques de NLP pas tout à fait convaincantes
- Représentation de la connaissance (information) extraite
  - Graphe connaissance
  - Rapprochement information multimodale
  - Analyse discours
- Restitution client / évaluation
  - Graphe + illustration par verbatim (pour générer « confiance » du client sur résultats obtenus)

Evaluation

- Des résultats d'analyse de polarité :
  - Une expérience d'évaluation participative
  - Réflexion sur la mise en place de « jeu » participatif et l'utilisation Amazon Turk
- De différentes techniques
  - Word embedding (word to vec : ex. Fastext https://fasttext.cc/) + Bi-LSTM (https://en.wikipedia.org/wiki/Long_short-term_memory)
  - Bert
  - Fine tuning ou transfert learning
- Analyse manuelle (human in the loop) des résultats obtenus par les réseaux de neurones :
  - Via une interface graphique dédiée : (Prodigy https://prodi.gy/) ou Watson Knowledge studio (www.ibm.com/fr-fr/cloud/watson-knowledge-studio)
- Interrogation sur les métriques mathématiques d'évaluation ; méthodes classiques en RI (F-Score, rappel, précision) vs métriques issues de la théorie de l'information (perplexité http://www-prima.inrialpes.fr/Vaufreydaz/These/Experimentations.html) ou Hamming Loss (https://hal.archives-ouvertes.fr/hal-01044994/document))

- Les métriques mathématiques, celles utilisées en TAL, sont utiles en terme de communication à destination du client pour leur dimension objective ; Cependant c'est la pertinence des verbatim associés aux insights mis en évidence qui emporte l'adhésion des clients
- Utilisé comme premier niveau d'évaluation, permet de dimensionner l'évaluation manuelle postérieure

Pratiques d'évaluation ergonomique au BRGM : Vincent GODARD (BRGM)

- Vincent Godard, chargé de l'ergonomie, communication, 3D et design UI/UX
- Maquettes fonctionnelles → maquettes interactives (liens hypertexte entre les pages, densité des liens) permettant une évaluation de site Web réaliste.
- Sur le produit fini, le BRGM manque de temps ou d'argent sur ses projets pour faire une objective des sites plus objective et plus propre et d'envergure plus grande. Mais début de réalisation de tests utilisateurs parfois.
- A la place, analyse directe des sites sur les traces d'usage (statistiques sur les pages parcourues, les liens suivis,..) + et surtout extraction de retours sur les verbatims de la plateforme d'assistance client : on retrouve indirectement une analyse des sentiments, mais plus ciblée telle que la pratique Aktan (*Aspect based sentiment analysis*)
- Claire Cailleau précise qu'une problématique pour le BRGM est celle de savoir comment évaluer a priori la complexité de présentation du contenu sur un sujet sur les sites du BRGM : complexité en termes de niveau de langue mais aussi de concepts présentés.

Mesures classiques d'évaluation appliquées à la classification de tweets ou la recherche d'information : Anne-Lyse MINARD (LLL)

- Evaluations basées sur référence (gold standard) : le standard (gold) est basé sur la réalisation de la tâche par des annotateurs humains.
- Evaluation intrinsèque : performance du système à réaliser la tâche
- Evaluation extrinsèque : lorsque le système est inclus dans un système plus large
- Objet d'évaluation :
    - Typage, catégorisation
    - Frontières (pour les entités)
- Outils : matrice de confusion qui permet de faire différentes mesures
- Mesure : selon les tâches affectées au système, c'est l'une ou l'autre des mesures ci-dessous qui doit être ciblée pour en mesurer la performance
    - Rappel : quelle est la couverture du système (exhaustivité)
    - Précision
    - Accuracy : part des Vrais Positifs dans l'ensemble
    - F-mesure : moyenne harmonique des mesures de rappel et de précision
- Evaluation globale ou locale pour la classification multi-classe
    - Macro moyenne : la moyenne est pondérée par la taille de chaque classe. Si l'ensemble de test est représentatif de la tâche en termes de distribution des classes,

permet de quantifier un score représentatif du comportement moyen du système sur la tâche
    - Micro moyenne : moyenne par classe non pondérée par la taille de ces classiques : n'est pas sensible au fait que les classes sont déséquilibrées. Métrique plus véridique (discriminante) des performances du système, car elle permet d'évaluer la performance du système classe par classe

- Evaluation pour le « sequence labeling » (détection entités)
    - Frontière
        - Strict match
        - Relaxed match : : autorisation d'une incertitude, à définir soi-même (un élément max de différence, 70%…)
    - Catégorie
- Test de significativité
- Exemples de campagnes d'évaluation
    - Tweet mentionnant des effets secondaires de médicaments (smm4h) mesure de la performance à détecter la classe
    - Détection d'informations temporelles : EVENTI à EVALITA 2014 (Casalli et al) évalué dans le cadre d'un système de questions-réponses : combien de bonnes réponses le système était capable de donner (oui/non) sur la base de son travail de détection
- Evaluation recherche d'information
    - Basé sur le ranking
    - + précision à k : nombre de documents pertinents parmi les k premiers documents

PARTICIPANTS

Industriels :

- AKTAN : https://aktan.fr/
    - Eric Clairambault : responsable R&D
        - Activité : développement outil TAL (analyse grammaticale avec Unitext + approche probabiliste avec réseaux de neurones)
- ATOS :
    - Olivier Gracianne
        - Stage : détection évènementielle sur Twitter
- QUALNET :
    - Marcel Raulier
        - Développeur : anonymisation de données sur « déclarations d'incidents » en hôpitaux ; intérêt pour la détection évènementielle dur Twitter
- EDF :
    - Julien Kahn :
        - Activité : EDF R&D « facteurs humains » docteur en ergonomie : RI dans les documents réglementaires
- BRGM :
    - Cécile Gracianne :
        - Activité : analyste de données : machine learning : extraction information

Académiques :

- LIFAT – Tours
  - Jean-Yves Antoine
  - Denis Maurel
    - Activité : détection d'entités nommées ; fouille de texte scientifique : détection des relations prédicat-argument sur des textes relatant des expériences pour mettre en évidence les « petits résultats » de recherche
  - Caroline Pasquet :
    - Activité : post-doctorat sur les expressions polylexicales
  - Adam Lion-Bouton :
    - Activité : évaluation des mesures d'évaluation des coréférences
- LLL - Orléans
  - Emmanuel Schang
  - Lotfi XXX :
    - Activité : injonctives
  - Flora Badin
    - Activité : ingénieure d'étude TAL - corpus oraux
  - Hélène Flamein
    - Activité : perception des orléanais de leur ville (corpus ESLO)
  - Anne-Lyse Minard :
    - Activité : extraction d'information en domaine médical + système anonymisation
- LIMSI – Orsay : https://www.limsi.fr/fr/
  - Sharleyne Lefebvre
    - Activité : résolution de coréférence et détection entités nommées dans les dialogues de séries télé