P18-2 – IMDR

Juillet 2022

« Mise en œuvre des technologies de traitement automatique du langage pour l'interprétation d'un retour d'expérience en langage naturel »

Quantmetry

IMdR Institut pour la Maîtrise des Risques
Sûreté de Fonctionnement - Management - Cindyniques

Chef de Projet :
M. Philippe DE LAHARPE (SNCF)

L'Institut pour la Maîtrise des Risques tient à remercier :

- M. Philippe DE LAHARPE– SNCF, Chef de projet, pour avoir dirigé cette étude,
- Les sociétés qui ont souscrit à ce projet et leurs collaborateurs qui ont participé à sa réalisation :

➢ CEA	M. Didier BALESTRIERI M. Frédéric DOYEN Mme. Françoise VAUFREY
➢ DGAC	M. Yoni MALKA
➢ EDF	M. Sylvain MAHE
➢ GRTgaz	Mme. Leïla MARLE Mme. Amel BELOUNNAS
➢ IMDR	M. Gaël GBONSOU M. Clément JUDEK M. André LANNOY
➢ IRSN	M. Richard LAUNAY, M. Jean-Marie ROUSSEAU
➢ RATP	M. Vianney BORDEAU Mme. Valérie HANOKA M. Vincent VILLE
➢ SNCF	M. Philippe DE LAHARPE Mme. Luce LEFEUVRE Mme. Coralie RETEUNER

- La société QUANTMETRY représentée par MM. Martin LE LOC, Karl NEUBERGER, Pierre-Jean LARPIN et Léandre ADAM-CUVILLIER,
- M. André Lannoy animateur de la commission Produits de l'IMdR.
- M. Vincent VILLE et Vianney BORDEAU pour la fourniture de données d'incidents et leurs analyses des résultats
- M. Richard LAUNAY, M. Jean-Marie ROUSSEAU, M. Didier BALESTRIERI, M. Frédéric DOYEN et Mme. Françoise VAUFREY pour la fourniture de rapport CRES

NOTE DE PRESENTATION ET DE SYNTHESE

Le TAL dans les REX

Les retours d'expérience (REX) sont au cœur des processus visant à **améliorer le contrôle qualité, la performance, la sécurité et la sûreté d'installations industrielles à haut risque** (réseaux de transports, centrales de production, …), qui reposent sur des **systèmes complexes sensibles**.

La valorisation des REX est ainsi indispensable. Constitués principalement de données non structurées sous forme de texte, **le Traitement Automatique du Langage (TAL) constitue un levier important pour les exploiter, aussi bien pour les extraire et les mettre en qualité** (transcription audio, transcription d'image, traduction…) **que pour les analyser** (détection d'événements critiques, regroupements d'événements, moteur de recherche…).

Bien que **l'automatisation du traitement des REX représente un enjeu majeur**, **leur exploitation reste aujourd'hui principalement manuelle** faute de modèles d'apprentissage automatique suffisamment performants.

Le domaine du Traitement Automatique du Langage a connu cependant une **rupture technologique** en 2017, avec **l'apparition de l'architecture *Transformers***, dont les modèles ont surpassé les performances humaines sur de nombreuses tâches. Ces modèles s'avèrent significativement plus performants pour abstraire l'information d'un texte. **Cette étude vise ainsi à comprendre les perspectives apportées par ces modèles puis à les éprouver sur des cas d'usages REX identifiés afin d'observer s'ils constituent une piste intéressante pour automatiser davantage de tâches dans les REX.**

L'état de l'art du TAL

Une caractéristique majeure du TAL est la capacité des modèles à **représenter les mots, phrases ou documents par des vecteurs numériques afin de pouvoir appliquer des méthodes d'optimisation mathématique. En outre, l'ensemble de ces représentations projetées dans un espace vectoriel doit garder une cohérence sémantique**, c'est-à-dire que les mots, phrases ou documents ayant un sens proche doivent garder une représentation similaire.

Les modèles basés sur l'architecture *Transformers* ont significativement amélioré les représentations vectorielles des textes, fournissant des plongements de textes qui s'adaptent même au contexte. Ces nouvelles architectures introduisent également une nouvelle méthode d'apprentissage en TAL **: l'apprentissage par transfert**. Un modèle se décompose ainsi en deux blocs séquentiels **: un premier bloc généraliste appelé modèle de langue dont le rôle est de convertir le texte dans des représentations vectorielles pertinentes, et un second bloc spécialiste, qui dépend de la tâche cible (génération de texte, classification, régression, regroupement…), et prend en entrée les représentations vectorielles du texte**. Ce nouveau paradigme se démarque ainsi des anciennes méthodes en proposant deux phases d'entraînement. La première, appelée **phase de pré-entraînement**, consiste à entraîner le modèle de langue sur des tâches dites **auto-supervisées**, avec un besoin en données non annotées très important. Une fois le modèle de langue pré-entraîné, **nous pouvons l'associer à un bloc spécialiste en fonction de la tâche et le spécifier sur des données annotées**. Cette nouvelle architecture permet à la fois de **diminuer drastiquement le nombre de données annotées nécessaires pour adresser un cas d'usage, et également de limiter le nombre de modèles à maintenir**. En effet, un modèle de langue unique pourra être associé à n'importe quel bloc spécifique pour adresser tous types de cas d'usages.

Aujourd'hui, **l'usage des modèles *Transformers* s'est fortement démocratisé**, notamment grâce à la mise en open-source de nombreux modèles de langue pré-entraînés, ou de modèles spécialisés sur une tâche cible. Ainsi, il est possible de charger un modèle de langue pré-entraîné par un pair sur un large corpus (comme Wikipédia) et de le spécifier ensuite sur une tâche avec des données annotées. Il est même possible d'utiliser directement des modèles spécialisés sur une tâche (Analyse de sentiments, Résumé de texte).

Les limites persistantes de l'état de l'art

Bien que les Transformers constituent une avancée considérable pour le TAL, il reste encore certains défis à relever pour **transposer et pérenniser toutes ces récentes prouesses académiques dans des projets industriels fiables et maintenables**.

Une des premières limites des modèles Transformers se situe dans leur forte complexité. En effet, **les performances des nouveaux modèles état de l'art ont été accompagnées d'une croissance phénoménale de la taille des modèles**, passant de centaines de milliers de paramètres à des milliards de paramètres. Cette croissance limite ainsi leur pratique sur des infrastructures adaptées comportant des processeurs graphiques disponibles sur le cloud ou sur des puissants serveurs locaux. Aussi, **cette sur-paramétrisation des modèles rend les résultats moins interprétables, complexifiant l'adoption des modèles par les métiers**. En effet, **l'usage de modèles de TAL en REX doit rester au service des experts métiers et constitue ainsi une Intelligence Augmentée pour accompagner l'expert dans ses décisions**. L'interprétabilité des modèles demeure par conséquent essentielle. Certes, la disponibilité des modèles en open-source représente un accélérateur considérable pour adresser rapidement un cas d'usage, cependant **cette forte dépendance à la communauté scientifique traduit également une limite pour les cas d'usages de TAL en français**, qui bénéficient d'une communauté moins dense qu'en anglais. Ainsi, certains modèles sont disponibles en anglais mais pas en français. Aussi, l'architecture *Transformers* demeure encore peu mature. **Elle nécessite notamment un coût extrêmement élevé en mémoire**, avec une complexité quadratique en fonction du nombre de mots, **limitant son usage à des séquences textuelles d'environ 500 mots**. De nouveaux modèles apparaissent pour adresser ce problème, mais leur portée reste académique à l'heure actuelle et non disponible pour des usages industriels en français.

Impact sur les cas d'usages REX

Afin d'évaluer l'impact des modèles état de l'art appliqués au REX, nous avons éprouvé quatre cas d'usages couvrant un large périmètre d'applications. Nous avons ainsi exploré un cas d'usage supervisé de classification de rapports de situations critiques (ANA-1), un cas d'usage non-supervisé de regroupements de rapports de situation (ANA-2), un cas d'usage de détection de rapports de situations similaires (ANA-3) et un cas d'usage de résumé automatique (ANA-5). Pour ces cas d'usages, **nous avons cherché à comparer les performances de modèles simples, basés sur des approches Sac de mots, avec des modèles état de l'art comme CamemBERT, un modèle BERT entraîné sur un corpus français.** De manière générale, **cette étude met en évidence la difficulté existante pour évaluer un modèle sur des tâches de TAL complexes**. En effet, certaines tâches donnent lieu à une multitude de réponses possibles qui ne peuvent finalement qu'être évaluées par validation manuelle d'un expert métier (ANA-3, ANA-5). Aussi, une tâche a priori simple comme ANA-1, est finalement complexe à analyser au regard des spécificités présentes dans la donnée et les labels.

Les cas d'usages ANA-2 et ANA-3 sont les plus explicites pour comparer le modèle CamemBERT avec les modèles références simples. En effet, leur étude compare la faculté d'un modèle à représenter les rapports de situations dans un espace vectoriel. Après consultation d'experts métiers, nous concluons **que le modèle CamemBERT infère des représentations vectorielles de textes plus pertinentes**. En effet, les regroupements et les événements similaires ressortis par le modèle CamemBERT sont nettement plus performants et **auraient aidé les experts métier dans leur travail, contrairement aux modèles références**.

Pour le cas d'usage ANA-1, nous n'observons pas de différence notable sur les métriques usuelles de classification entre le modèle CamemBERT et le modèle TF-IDF. Certes, l'étude plus approfondie des erreurs nous montre que le modèle CamemBERT parvient à mieux traiter des subtilités comme la différence entre un objet tombé sur les voies et un objet partiellement coincé, toutefois les résultats d'ensemble sur ce cas d'usage restent décevants au regard des différences observées sur les cas d'usage ANA-2 et ANA-3. Une brève étude a permis d'observer une proportion non négligeable de labels non pertinents, expliquant potentiellement la difficulté existante pour analyser ce cas d'usage.

Enfin, le cas d'usage ANA-5 présente des résultats non concluants. Ne disposant pas de données labellisées, nous avons éprouvé un modèle de résumé automatique entraîné à générer la synthèse

d'articles de journaux et disponible en open-source. Bien que fournissant une synthèse intelligible, le résumé proposé par le modèle ne parvenait pas à extraire l'information pertinente du rapport**, car les rapports CRES présentent une structure trop éloignée d'articles journalistiques et requièrent une expertise pointue du métier pour isoler l'information pertinente. De plus, les rapports CRES sont très longs et se confrontent donc aux limites des Transformers aux textes longs.**

Globalement, les résultats apportés par les modèles état de l'art sont prometteurs et méritent d'être approfondis, bien qu'il soit difficile d'en tirer des conclusions définitives dans un contexte d'application aussi bref. Aussi, nous observons **une limite importante quant à l'application de modèles état de l'art entraînés sur des corpus généralistes pour un domaine métier aussi spécifique**. Il semble nécessaire d'envisager une spécification de ces modèles sur un corpus spécifique pour davantage tirer profit de ces modèles.

La feuille de route d'un cas d'usage TAL en REX

1. **Définir clairement le processus d'évaluation correspondant aux attentes du métier**. Dans l'idéal, il est souhaitable de constituer un jeu d'évaluation référence en amont qui permettra de mesurer la performance du modèle par rapport aux attentes métiers.
2. **Partir d'une solution simple comme point de référence (modèle Sac de mots)**
3. **Selon la difficulté du cas d'usage et la performance attendue par les métiers, tester les modèles Transformers comme CamemBERT, en partant de son application la plus simple.**
    a. Application d'un modèle disponible en open-source et déjà spécifié sur la tâche cible (comme ANA-5)
    b. Utilisation d'un modèle de langue pré-entraîné sur un corpus généraliste et spécifier sur la tâche cible avec un jeu de données annotées métier comme pour ANA-1 (des centaines de données suffisent pour une première itération, même si cela dépend du cas d'usage)
    c. Réentraînement d'un modèle de langue sur un corpus métier. Cette dernière phase nécessite des infrastructures dédiées avec des GPU et un volume élevé de données métiers (plusieurs millions de phrases, variables selon un réentraînement complet ou partiel). Une fois cette phase terminée, le nouveau modèle de langue peut être utilisé dans tous les cas d'usages.

Enfin, les cas d'usages éprouvés ont mis en valeur l'importance **d'avoir une cohabitation étroite avec les experts métiers.** Il est indispensable de mettre en place des processus efficaces pour favoriser les interactions avec les experts métiers. Ces processus se matérialisent par l'implémentation d'outils d'évaluation ergonomiques et des interactions récurrentes pour échanger sur les dernières avancées.

Perspectives

Essentiellement focalisé sur l'apport des modèles état de l'art, le projet IMdR P18-2 peut être complété au travers de plusieurs voies d'étude ou de recherche pour étudier plus en profondeur les autres barrières existantes à la mise en œuvre de projets TAL :

- Étude sur l'extraction et la mise en qualité de données non-structurées (Est-ce un facteur bloquant pour développer des cas d'usages TAL, quel impact sur la gouvernance de la donnée ?)
- Définition et mise en place de processus standards et ergonomiques d'annotations de données

Dans la continuité de l'étude, le projet IMDR P18-2 peut également se prolonger sur plusieurs voies d'étude ou de recherche telles que :

- Étude de l'impact d'un modèle Transformers spécifié sur un domaine métier (qu'est-ce qu'un domaine métier ? Mesurer l'impact sur la qualité des plongements lexicaux)
- Étude de l'interprétabilité des modèles Transformers dans une perspective d'Intelligence Augmentée