Apprentissage de représentation d'opérateurs d'extraction de contenu pour des applications en data-to-text

Encadrants : Sylvain Lamprier (SU) - Sylvain.Lamprier@lip6.fr
Benjamin Piwowarski (SU) - Benjamin.Piwowarski@lip6.fr

Lieu du Stage : Jussieu
Durée : 6 mois
Compétences souhaitées : de bonnes compétences en apprentissage par renforcement profond sont requises. Un bon niveau de programmation et la maîtrise d'une librairie de *deep learning* (pyTorch de préférence) sont également demandés.

Mots-clé : Data-to-Text; Apprentissage de représentations; Réseaux Transformers; Extraction de contenu

Contexte

La très grande disponibilité des données est un fait bien établi dans notre société. Que les données proviennent de textes, de traces d'utilisateurs, de capteurs ou encore de bases de connaissances, l'un des défis communs est de comprendre et d'accéder rapidement aux informations contenues dans ces données pour faciliter la prise de décision. Une des réponses à ce défi consiste à générer des synthèses textuelles des données considérées, le langage naturel présentant de nombreux avantages en termes d'interprétabilité, de compositionnalité, d'accessibilité et de transférabilité. Néanmoins, si la génération de résumés pour données textuelles est un problème pour lequel les solutions commencent à être satisfaisantes, la génération de descriptions textuelles dans un cadre plus général (e.g., conditionnelles à des données numériques ou structurées) constitue toujours un problème particulièrement difficile. Ce problème fait référence à un champ émergent dans le domaine du traitement du langage naturel, appelé Data-to-Text, possédant de très nombreuses applications, notamment dans les domaines scientifiques, du journalisme, de la santé, du marketing, de la finance, etc. Un des premiers exemples d'application fut la publication d'un article du Los Angeles Time, généré automatiquement à partir de données numériques sismiques. D'autres exemples ont concerné le suivi des flux numériques (bourse, billetterie, suivi de la population, etc.), l'assistance aux diagnostics médicaux ou encore le soutien d'enfants en difficulté d'élocution (par exemple, pour les aider à mieux retranscrire leurs journées). Une agence qui analyse des rapports d'entreprises pour simuler des stress tests écologiques sur des milliers de produits financiers, nous a rapporté que l'information utile de ces rapports se situait à 60% dans des tableaux, 10% dans des graphiques et seulement 30% dans le texte des rapports. Cet exemple illustre l'importance du problème, que les avancées récentes en apprentissage profond et génération de la langue (e.g., via des

réseaux type BERT, GPT, etc.), rendent possible à envisager. Le projet ANR ACDC dont le démarrage est prévu pour début Avril 2022, et dont l'équipe MLIA est coordinatrice, s'appuie sur ces avancées pour la génération de synthèses textuelles à partir de données tabulaires (bien que les propositions pourraient ultérieurement être étendues à d'autres types de données structurées telles des séries numériques, figures ou graphes), avec un accent particulier porté sur la recherche d'invariance des données d'entrée, l'extraction d'opérateurs de sélection/compression haut-niveau et la personnalisation des sorties produites.

L'ensemble des approches récentes de data-to-text travaillent de manière supervisée, sans représentation explicite des opérateurs d'extraction qu'ils manipulent pour passer du contenu tabulaire global à la synthèse textuelle [1,2,3,4]. Ce projet se démarque car il propose de s'intéresser à l'expression de ces opérateurs, afin de gagner en interprétabilité des modèles, ainsi qu'en capacité de contrôle sur les textes générés. En outre, si dans un cadre figé bien défini, avec de nombreuses ressources pour la supervision, il est possible de s'affranchir de l'expression explicite de ces opérateurs, car le mode de sélection peut être implicitement adapté en fonction des sorties désirées, ce n'est plus envisageable dans un cadre plus large avec une grande hétérogénéité des données d'entrée et des attendus dans un contexte où la supervision est limitée. Notre démarche, en forte rupture avec les approches de la littérature, est donc de chercher à inférer les opérateurs d'extraction de contenu permettant de passer d'un tableau à un texte observé, en ayant pour but d'avoir un apprentissage robuste, qui soit à la fois fortement généralisable et contrôlable par un utilisateur.

Pour répondre à ces besoins, nous proposons la construction d'un espace latent sémantique des opérateurs sur les tableaux. Celui-ci correspond à l'aspect novateur majeur sur lequel s'organise le projet. Une propriété désirée pour cet espace, que l'on cherchera à satisfaire au cours du projet, correspond au fait que de tout opérateur doit être adaptable à tout tableau sur lequel on souhaite l'appliquer. D'un autre côté, on souhaite qu'un maximum de sémantique soit préservée d'un tableau à l'autre pour un même opérateur, afin de disposer d'un espace dans lequel il est pratique d'apprendre des stratégies d'échantillonnage des opérateurs, dont les éléments ont un effet similaire quel que soit le contexte dans lequel ils sont appliqués. On ne souhaite alors pas que les opérateurs, qui correspondent initialement à des requêtes de type SQL par exemple, soient encodés en absolu dans l'espace de représentation Ξ visé, mais plutôt qu'elles soient exprimées selon des proximités dans un référentiel sémantique. Par exemple, on souhaite qu'une opération, correspondant à réaliser la moyenne du poids de tous les orang-outan d'un tableau de poids d'animaux, soit encodée comme devant retourner la moyenne du poids de l'espèce la plus proche de l'orang-outan du tableau sur lequel on l'applique, selon un espace sémantique à définir. De la même manière sur les noms des attributs plutôt que leur valeur, on souhaite qu'un opérateur sur des colonnes de poids puisse s'appliquer sur des colonnes sans nom explicite d'attribut mais dont les valeurs sont exprimées en kg (et inversement). Enfin, on souhaite exprimer une certaine proximité entre les types d'opérations d'extraction utilisés.

Cette proposition, qui correspond à un encodage contextualisable des opérateurs d'extraction, permettra à notre espace latent d'être robuste à l'hétérogénéité des tableaux, et

d'être efficace lors du transfert du modèle à d'autres données. Lors du décodage d'un opérateur, il est alors toujours possible d'extraire une opération valide, dont nous pourrons apprendre des distributions de probabilités en fonction des tableaux considérés et des textes descriptifs visés. En outre, un espace présentant de telles propriétés permet de considérablement simplifier les mécanismes d'inférence d'opérateurs à partir de couples tableau-texte, car une correspondance directe entre les opérateurs et le texte peut être trouvée (i.e., $p(\xi|\tau, \omega) \approx p(\xi|\omega)$, avec τ un tableau et ω la description textuelle correspondante). Afin de tendre vers ce genre d'espace contextualisable, nous définirons divers mécanismes d'encodage-décodage avec coûts adverses, basés sur des architectures neuronales avec réseaux d'attention du type Transformer Network, et des transformations contre-factuelles des tableaux considérés. Des politiques de combinaisons d'opérateurs, séquentielles ou en parallèles, dans cet espace seront également considérées, afin d'élargir le spectre des synthèses de tableau possibles, en évitant l'explosion de la complexité de l'espace de représentation. De tels travaux, traitant à la fois de l'hétérogénéité (i.e., variabilité des entrées et des attendus) et de la recherche d'opérateurs d'extraction pour la génération textuelle, sont tout à fait novateurs dans le domaine du data-to-text, et même au-delà, pour l'extraction de contenu synthétique de manière plus générale.

Description

Dans ce stage, nous considérons qu'une unité textuelle de description (e.g., une phrase) correspond à une vue sur un tableau global, que l'on peut représenter de manière équivalente par une structure tabulaire associée à une opération algébrique d'extraction. La stage s'intéresse à l'apprentissage de représentations sémantiques ξ dans un espace Ξ des opérations d'extraction dans les tableaux. L'objectif est d'encoder des opérateurs contextualisés, i.e. pour lesquels il est possible d'adapter les effets aux tableaux d'entrée considérés relativement à leur contenu/structure. L'espace de représentation appris servira de base aux tâches suivantes du projet ACDC, pour 1) l'inférence de descriptions textuelles via l'apprentissage de la distribution de probabilités d'opérations élémentaires de notre espace appris, et 2) la composition d'opérateurs pour des inférences textuelles plus complexes.

À partir d'un générateur d'expressions algébriques dans un format prédéfini (un sous-ensemble de l'algèbre des requêtes SQL) permettant de générer des vues sur des tableaux, nous chercherons donc au cours de ce stage à apprendre un espace d'encodage sémantique des opérateurs sur les tableaux de manière auto-supervisée. Divers types d'opérateurs élémentaires, composés d'opérations algébriques unaires (e.g., sélection, projection, etc.) et d'agrégation (e.g., maximum, moyenne, comptage, fréquence, etc.) pourront être considérés. Nous définirons ainsi une algèbre de transformation dans notre espace, que nous pourrons adapter pour modifier la complexité et l'expressivité de nos modèles.

Plus précisément, on considérera des couples (τ, s), avec $\tau \in \Gamma$ un tableau d'entités et $s \in S$ une opération élémentaire de l'algèbre telle que s est adaptée à τ. On cherchera une

fonction d'encodage d'opérateurs (incluant leurs arguments) e : Γ×S → Ξ et une fonction de décodage pour un tableau d'entrée d : Γ×Ξ → S̃, ainsi qu'une fonction d'interprétation f : Γ × S̃ → Γ, telles que f(τ, ŝ) ≈ s(τ), avec ŝ = d(τ, e(τ, s)) l'opérateur adapté à τ dans un espace de décodage S̃, et s(τ) l'application de l'opération s à τ selon l'interpréteur à disposition pendant cette phase auto-supervisée. L'idée est de comparer ŝ à l'opérateur original s, selon son effet sur τ via la fonction f, qui vise à mimer l'interpréteur selon des éléments de S̃. Cette approche, flexible car ne nécessitant pas de décoder des opérations syntaxiquement valides pour l'interpréteur, permet de définir des coûts dérivables dans l'espace des tableaux d'arrivée. Pour les différentes fonctions, on envisage d'utiliser des réseaux transformeurs neuronaux avec encodage de structure (cellule, colonne et ligne) s'inspirant de ceux employés sur des graphes (e.g., GTNs [5]). La comparaison des sorties tabulaires se fera par des métriques invariantes aux permutations.

Un objectif majeur est que l'encodeur utilise le contenu du tableau qui lui est passé en entrée pour encoder l'opérateur dans un espace Ξ, et respecte la sémantique de la requête relativement aux éléments du tableau, plutôt que d'encoder les arguments de manière absolue. Le décodeur vise à adapter tout opérateur issu de l'espace Ξ au tableau sur lequel on souhaite l'appliquer. Pour atteindre ces objectifs, le projet prévoit de considérer des coûts adverses qui tendent à modifier les paramètres de l'encodeur de manière à ce que le décodage selon un tableau contre-factuel de sémantique différente τ̂ produise un opérateur d'effet bien différent sur le tableau original (i.e., ŝ(τ) très différent de s(τ) pour ŝ = d(τ̂ , e(τ, s), avec ŝ(τ) l'application de l'opérateur décodé au tableau τ selon f). Cela forcera l'encodeur à se servir de τ . Pour obtenir une sémantique haut-niveau, il faut cependant que l'on évite d'encoder les arguments des opérateurs selon des indicateurs pauvres, du type indices des lignes et colonnes qu'ils manipulent. On propose alors d'encoder les opérateurs en appliquant des perturbations à τ, telles que des permutations de ses lignes et colonnes (voire des modifications de τ selon des symétries sémantiques).

Le stage s'articulera autour des étapes suivantes:

1. Identification des ressources tabulaires à utiliser et des opérations à encoder sur ces tableaux (ceci pourra se faire selon des jeux de données académiques du domaine ou bien des données issues de notre partenaire Muséum National d'Histoire Naturel);
2. Production d'un générateur artificiel d'opérations élémentaires en SQL (via automates);
3. Définition de métriques de comparaison de tableaux avec recherche d'invariance aux permutations structurelles ou symétries sémantiques;
4. Encodage d'opérateurs de sélection et projection, avec recherche d'invariances structurelles et sémantiques;
5. Expérimentations selon diverses métriques d'analyse de l'espace appris;
6. Encodage d'opérateurs d'agrégation et expérimentations;
7. Si le temps le permet: Inférence d'opérateurs de l'espace à partir de textes (soit de manière supervisée selon des textes annotés requêtes, soit de manière non-supervisé par auto-encodage variationnel des textes considérés). L'objectif est, pour chaque texte d'entrée considéré, d'être à même d'identifier la vue qui l'a engendré. Cela a trait au domaine très actif du Semantic Parsing [6], mais en se

focalisant sur le résultat de la requête, plutôt que sa simple traduction en langage naturel.

Références

1. S. Agarwal and M. Dymetman. “A surprisingly effective out-of-the-box char2char model on the E2E NLG Challenge dataset”. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. 2017.
2. R. Lebret, D. Grangier, and M. Auli. “Neural Text Generation from Structured Data with Application to the Biography Domain”. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.
3. S. Wiseman, S. Shieber, and A. Rush. “Challenges in Data-to-Document Generation”. In: Empirical Methods in Natural Language Processing. 2017.
4. C. Rebuffel, L. Soulier, G. Scoutheeten, and P. Gallinari. “A Hierarchical Model for Data-to-Text Generation”. In: ECIR 2020.
5. S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim. “Graph Transformer Networks”. In: CoRR abs/1911.06455 (2019).
6. B. Wang, M. Lapata, and I. Titov. “Learning from Executions for Semantic Parsing”. In: arXiv:2104.05819 (2021).