INSTITUT DES SYSTEMES INTELLIGENTS ET DE ROBOTIQUE
OFFRE DE THÈSE

Sujet de thèse

Titre de la thèse : Traduction automatique de documents: le cas des documents scientifiques

Directrice ou directeur de thèse : François Yvon

Co-direction éventuelle : Rachel Bawden (Inria Paris)

Laboratoire d'accueil : ISIR (*Institut des Systèmes Intelligents et de Robotique*), Campus Pierre et Marie Curie, 4 place Jussieu, 75005 Paris.

Personne à contacter

Prénom Nom : François YVON

Email : yvon(at)isir.upmc.fr

Envoyer votre candidature par mail, avec [*sujet de la thèse*] en objet, un CV et une lettre de motivation.

Description du sujet (en français)

Contexte :
La thèse se déroule dans le cadre du projet ANR-MaTOS (https://www.anr-matos.fr). Le candidat sera co-encadré par François Yvon (DR CNRS) au sein de l'équipe MLIA de l'Institut des Systèmes Intelligents et de Robotique (https://isir.upmc.fr) et Rachel Bawden, CR Inria, dans l'équipe Almanach du Centre Inria Paris (https://www.inria.fr/fr/centre-inria-de-paris) et inscrit dans [l'école doctorale informatique, télécommunication et électronique (EDITE)](https://www.edite-de-paris.fr/) de Paris.

Le ou la candidate recruté-e bénéficiera d'un contrat doctoral de Sorbonne Université, et aura, si il ou elle le souhaite, l'opportunité de réaliser des missions d'enseignement au sein de l'Université.

Description du projet :
La plupart des systèmes de traduction automatique (TA) dits "neuronaux" modélisent le processus de génération d'un document cible y à partir d'un document source x en décomposant ce processus phrase par phrase, chaque phrase étant traduite indépendamment des phrases voisines. Le modèle probabiliste sous-jacent est alors de la forme $P(y|x; \theta)$, où θ représente l'ensemble des paramètres du modèle (par exemple un modèle Transformer [Vas17]). Une fois θ appris, la génération d'une traduction repose sur la recherche de la traduction la plus probable réalisant: $\arg \max_{y} P(y|x; \theta)$.

Cette modélisation est naïve et ignore les multiples dépendances qui existent entre les phrases au sein d'un même document. Pour pallier cette déficience, de nombreuses architectures alternatives ont été proposées pour intégrer un contexte discursif c au modèle, conduisant à des modèles de la forme $P(y|x, c; \theta)$. Selon les implantations, le contexte c représente quelques phrases précédent x, ou bien tout le document source, ou bien également le début de la traduction (les phrases précédant y). Plusieurs manières d'encoder c (avec un encodeur dédié, ou bien en utilisant le même encodeur que pour x) ont été proposées dans la litérature. Les principales architectures de ce type, dédiées à la TA de documents (TAD), sont décrites dans [Mar21].

Deux obstacles principaux rendent cette extension difficile à implanter : (a) les ressources computationnelles (mémoire et calcul) nécesssaires à l'encodage d'un contexte étendu croissent de manière quadratique avec la longueur du contexte (pour les architectures Transformer); (b) l'apprentissage des dépendances entre y et c est rendu difficile par la relative rareté des mots pour lesquels le contexte étendu c est utile. La plupart des études dans le cadre de la TAD s'intéressent au problème (a) et considèrent soit des approximations du calcul de l'attention (voir [Tay21] pour un état des lieux récent), soit des architectures alternatives au modèle Transformer (par exemple [Gu21]) pour encoder des séquences.

Objectif scientifique :
Le travail de thèse proposé s'intéresse au problème de la traduction de documents complets, en se focalisant sur un type de documents particulier, à savoir les écrits académiques (articles, communications, rapports de recherche). Il s'agit de documents relativement longs, qui sont régis par des principes d'organisation et de présentation rigides et propres à ce genre textuel - organisation en sections en sous-sections -, ainsi que par des stratégies argumentatives propres à ce genre de textes: introduction de concepts et de définititions, raisonnements explicites devant venir à l'appui de démonstrations et de conclusions précises, etc.

L'objectif principal de la thèse est de parvenir à assurer que les documents générés par traduction automatique (a) reproduisent correctement la structure générale du texte d'entrée; (b) manifestent le même niveau de cohésion et de cohérence, en particulier dans le choix des termes, que le texte source; (c) reproduisent fidèlement les stratégies argumentatives (prémisses, déductions, conclusions) qui sont présentes dans le texte source, et (d) énoncent dans la langue cible les mêmes conclusions générales que dans la langue source. Parmi les autres difficultés de la tâche, qui pourront faire l'objet d'une attention particulière, mentionnons: la présence de nombreux extra-lexicaux (chiffres, symboles mathématiques, noms propres) et de parties non-textuelles (formules, équations, tableaux, graphiques). Notons enfin que les méthodes considérées devront être adaptées à une situation où les données monolingues sont abondantes, mais les données parallèles sont extrêmement rares: ce contexte est propice à l'utilisation de grands modèles de langue pré-entrainés.

Pour parvenir à ces fins, on s'intéressera par exemple aux architectures, déjà évoquées ci-dessus, qui exploitent un contexte discursif étendu, que ce soit pour la traduction automatique ou pour le résumé de longs documents [Koh22], ou encore aux méthodes de planification également utilisées pour la génération automatique de textes [Pup22]. L'enjeu principal de cet axe sera de rechercher les meilleurs compromis entre la complexité algorithmique du traitement de grands contextes (à l'apprentissage et à l'inférence) et le bénéfice tangible de ces efforts mesuré par de l'accomplissement des objectifs (a-d).

Un second axe de travail s'intéressera à la modélisation des stratégies discursives et des objectifs de communication associés à chacune des parties du document: une manière simple d'approximer ces objectifs s'appuie sur la structure interne des documents, mais des approches plus sophistiquées, utilisant des modèles à données latentes, devront également être considérées.

Profil recherché :
Master en Informatique ou Mathématiques Appliquées avec une spécialisation en Intelligence Artificielle, Apprentisage Automatique, Traitement des Langues, ou diplôme équivalent).

Compétences requises :
- Solides compétences en programmation (PyTorch),

- Communication en anglais écrit et parlé
- Créativité et capacité à formuler et à résoudre des problèmes de manière autonome

Description du sujet (en anglais)

Title : Document-level Machine translation: Translating Scientific Texts

Context:
This PhD will be fully founded by the ANR project MaTOS Machine Translation for Open Science(https://www.anr-matos.fr) which aims to develop new methods of automatically translating and evaluating scientific documents. The project focuses on translation between English and French, for which resources are readily available and translations are of a reasonable quality and coherence. The PhD will be co-supervised by Rachel Bawden (Inria) and François Yvon (CNRS).

Project Description:
Most so-called "neural" machine translation (MT) systems model the process of generating a target language document y from a source language document x by decomposing this process sentence by sentence, each sentence being translated independently of the neighboring sentences. The underlying probabilistic model takes thus the following form: $P(y|x; \theta)$, where θ represents the set of parameters of the model (for example a Transformer model [Vas17]). Once θ is learned, the generation of a translation is based on the search for the most probable translation realizing: $\arg \max_{y} P(y|x; \theta)$.

Such models are naive and ignore the multiple dependencies that exist between sentences within the same document. To overcome this deficiency, multiple alternative architectures have been proposed to integrate a discourse context c into the model, leading to models of the form $P(y|x, c; \theta)$. Depending on the implementation, the context c represents a few sentences before x, or the whole source document, or the beginning of the translation (the sentences before y). Several ways of encoding c (with a dedicated encoder, or using the same encoder as for x) have been proposed in the literature. The most common such architectures, dedicated to document MT (DLTM), are described in [Mar21].

Two main obstacles make this extension difficult to implement: (a) the computational resources (memory and computation time) required to encode an extended context grow quadratically with the length of the context (for Transformer architectures); (b) learning the dependencies between y and c is made difficult by the relative scarcity of words for which the extended context c is useful. Most studies in the DLMT framework address problem (a) and consider either approximations to the attention computation (see [Tay21] for a recent review) or alternative architectures to the Transformer model (e.g. [Gu21]) for encoding long sequences.

This thesis proposal addresses the problem of translating complete documents, focusing on a particular type of document: academic papers (articles, communications, research reports). These documents are relatively long, and are governed by rigid principles of organization and presentation specific to this genre of texts - division into sections and subsections - as well as by specific argumentative strategies: introduction of concepts and definitions, explicit reasoning to support precise demonstrations and conclusions, etc.

Scientific Objective:
The main objectives of the thesis are to ensure that the documents generated by machine translation (a) correctly reproduce the general structure of the input text; (b) display the same level of cohesion and coherence, especially in the choice of terms, as the source text; (c) faithfully

reproduce the argumentative strategies (premises, deductions, conclusions) that are present in the source text, and (d) state the same general conclusions in the target language as in the source language. Other difficulties of the task, which may require special attention, include: the presence of many extra-lexicals (numbers, mathematical symbols, proper names) and non-textual parts (formulas, equations, tables, graphs). Finally, it should be noted that the methods considered will have to be adapted to a situation where monolingual data are abundant, but parallel data are extremely rare: this context suggests to consider the use of large pre-trained language models.

To achieve these goals, we will be interested for instance in architectures that exploit an extended discourse context (see references below) whether proposed for machine translation or for long document summarization [Koh22]. We will also consider planification methods that are used for automatic text generation [Pup22]. The main challenge of this line of work will be to find the best trade-offs between the algorithmic complexity of processing large contexts (for learning and inference) and the tangible benefit of these efforts as measured by the achievement of objectives (a-d).

A second line of work will focus on modelling the discourse strategies and communication goals associated with each part of the document: a simple way of approximating these goals is based on the internal structure of the documents, but more sophisticated approaches, using latent variable models, will also have to be considered.

Required Profile:
Candidates should have a Master 2 or equivalent (e.g. engineering school) in computer science (speciality artificial intelligence, machine learning or natural language processing).

Required skills:
The candidate should have a good level in programming (python), experience with neural networks and an interest in natural language processing. A good written and spoken level of English is required, and knowledge of French is preferred.
We are looking for highly motivated candidates with a strong background in NLP, machine learning and an interest in linguistics and language. Ideally, candidates should be able to show initiative, creativity and have a good eye for analysis of data and results.

References:
* [Gu21] A. Gu, K. Goel, and C. Ré. Efficiently modeling long sequences with structured state spaces. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.
* [Koh22] H. Y. Koh, J. Ju, M. Liu, and S. Pan. An empirical survey on long document summarization: Datasets, models, and metrics. ACM Comput. Surv., 55(8), dec 2022.
* [Mar21] S. Maruf, F. Saleh, and G. Haffari. A survey on document-level neural machine translation: Methods and evaluation. ACM Comput. Surv., 54(2), Mar. 2021.
* [Pup22] R. Puduppully, Y. Fu, and M. Lapata. Data-to-text generation with variational sequential planning. Transactions of the Association for Computational Linguistics, 10:697–715, 2022.
* [Tay21] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler. Efficient transformers: A survey. ACM Comput. Surv., apr 2022.
* [Vas17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc., 2017.