Document-level Machine translation: Translating Scientific Texts

François Yvon - Rachel Bawden

April 2023

Scientific Project

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 neighbouring 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 organisation 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.

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.

Context This PhD will be financed by the ANR project MaTOS Machine Translation for Open Science 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).

Main activities

The main activities of the PhD will include :

— keeping up-to-date with related work on the topic, and producing a report on the state of research in the field in the context of the ANR project

- carrying out research on the topic outlined above, both in the development of new ideas, positioning with respect to related work and validation of the methodology via experiments and analysis
- the presentation of work both internally to colleagues and externally in the form of conference/journal/workshop papers and in the final PhD thesis
- interacting and exchanging with colleagues on NLP topics

Expertise required and qualities sought

The position is a 3-year funded PhD position of starting on the 1st September 2023 at the earliest. Candidates should have a Master 2 or equivalent (e.g. engineering school) in computer science (speciality artificial intelligence, machine learning or natural language processing). They 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. In your application (which can be in English or in French), please include :

- Curicullum Vitae
- Letter of motivation
- Letters of recommendation
- Optionally an example of your previous written work (if possible related to NLP), for example a master's thesis, research paper, etc.

Apply by email to François Yvon (first.last@cnrs.fr) and Rachel Bawden (first.last@inria.fr) with subject "[Phd-DLMT] Application of First Last" *before the 22nd of may 2023*.

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.