Conversational Fine-Tuning of italian T5 model

Simone Caldarella
simone.caldarella@studenti.unitn.it

1 Introduction

This is the report for the internal internship carried out from February 1^{st} to April 30^{th} (150h - 6 CFU) at SISLab (Signal and Interactive System Lab) under the supervision of Professor G. Riccardi and co-supervised by PhD Student (S.) Mahed Mousavi.

The following report is divided into four sections and, apart of the first that will explain the background and the general topic behind the internship, each of the other three is related to one milestone reached, in term of model/technique developed, and contains description of the work done, papers used to deep dive in the topic and results obtained along with other achievements.

2 Knowledge Grounded Dialogue Response Generation on FOUPs dataset

The main line of the internship was to build and fine-tune deep models (along with related analysis) on the FOUPs dataset (Mousavi et al. (2021)).

The goal was to build an End-to-End Model that could learn how to hold a therapeutic dialogue in a choerent and contextualized way. One of the key points of the project involved the investigation and implementation of grounding models for dialogue generation, which refers to models that the can produce answers that are implicitly or explicitly linked to the dialogue history and the knowledge of what has happened before the actual dialogue took place (named "context" in the dataset). And example of this could be:

- Context: "Sono andato al lavoro e ho litigato con un collega, il quale mi ha urlato contro"
- Dialogue History:
  - PHA: "Hai scritto di essere andato al lavoro e di aver litigato con un collega. Cosa è successo?"
  - Patient: "Sono arrivato al lavoro, ma non vedendo il collega non l'ho salutato. A lui è sembrata una mancanza di rispetto e mi ha urlato contro"
- Model Response: "Hai provato a parlarne con *lui* di come ti senti a riguardo?"

Moreover, it was investigated how different ways of providing knowledge could affect the groundness of the response. The different knwoledge pieces were extracted using automatic and unsupervised methodologies and were of three different kinds:

- Raw, which refers to the knowledge provided as it is, in an unprocessed text format;
- Bag of Words (BoW), which refers to the list of the extracted most relevant words from a knowledge piece (mostly nouns and verbs);
- Personal Space Graph (PSG), which refers to a graph structure where the raw knowledge is parsed and processed in a graph structure that keep track of event relations between partecipants (Mousavi et al. (2022)).

A second key point was the evaluation of the model responses, in order to understand, at a different level, the goodness of a response (appropriatness, correctness, contextualization and sequentiality).

Following the literature, the performance of dialogue generation models were assessed employing both automatic evaluation metrics (BLEU, ROUGE, etc.) and human evaluation metrics. In this step, the novelty part lied in the evaluated dimensions, which include=coeherentness and contextualization, in order to accomplish a deeper assessment of the model behaviour.

3 Recurrent Sequence-to-Sequence baseline model

As a first milestone of the internship, I implemented a Recurrent Sequence to Sequence model, starting from the work done by Sutskever et al. (2014).

The first step of this milestone, after having parsed all the necessary data, was to compute the vocabulary coverage between all the italians word embeddings available and the vocabulary extracted by FOUPs training set and rank them.

The second step was to develop the sequence to sequence model (called Seq2Seq) and evaluate it using the following metrics:

- Negative Log Likelihood;
- Perplexity;
- BLEU-2;
- BLEU-4;

The model, sketched below (Fig. 1), consist of an encoder part, which encodes the original embedded input text into a fixed length vector, and a decoder part, which takes as input the vector generated by the encoder and tries to generate the corresponding output sequence.

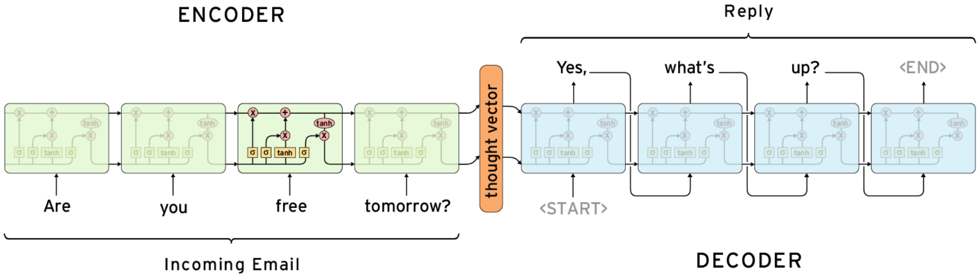


Figure 1: Recurrent Sequence to Sequence model.

Both the encoder and the decoder consist of a recurrent neural network (RNN, LSTM or GRU) which sequentially encodes (/generate) the input (/output) word by word. This sequential behavior of this kind of models permits to deal with variable length inputs and outputs. However, there is a well known problem using this architecture: encoding a variable length input, with both short and long interconnections, into a fixed length (and usually small) vector leads to an high loss of information. For this reason, in order to not loose relevant informations between encoder and decoder, the attention mechanism was introduced by Bahdanau et al. (2015). Originally developed for neural machine translation tasks, suddenly became the standard "solution" for sequence to sequence (and not only) tasks. The salient point of attention mechanism lies in the jointed-learned *attention weights* used in the weighted sum of the input in order to obtain an information flow that gives more importance to the predicted more relevant features of the input (Fig. 2).

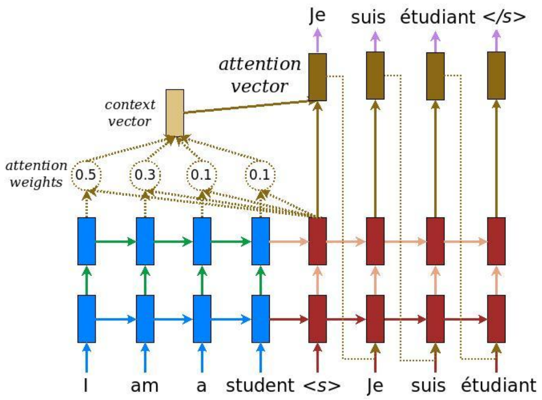
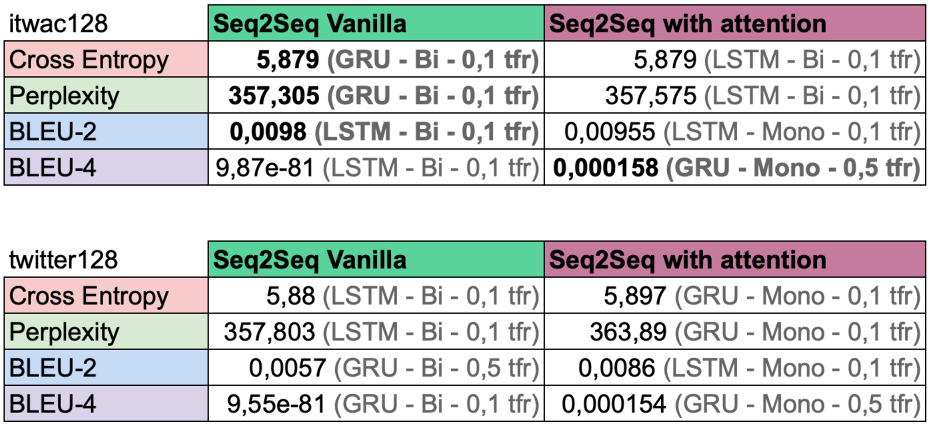


Figure 2: Sequence to Sequence with Attention Mechanism.

Once having developed the two variants of Recurrent sequence to Sequence model and having performed a shallow hyperparameters tuning, I proceeded with the automatic evaluation phase, whose results (Fig. 3 revealed the inadequacy of this architecture to deal with our small dataset in a training-from-scratch setting.



itwac128	Seq2Seq Vanilla	Seq2Seq with attention
Cross Entropy	**5,879 (GRU - Bi - 0,1 tfr)**	5,879 (LSTM - Bi - 0,1 tfr)
Perplexity	**357,305 (GRU - Bi - 0,1 tfr)**	357,575 (LSTM - Bi - 0,1 tfr)
BLEU-2	**0,0098 (LSTM - Bi - 0,1 tfr)**	0,00955 (LSTM - Mono - 0,1 tfr)
BLEU-4	9,87e-81 (LSTM - Bi - 0,1 tfr)	**0,000158 (GRU - Mono - 0,5 tfr)**

twitter128	Seq2Seq Vanilla	Seq2Seq with attention
Cross Entropy	5,88 (LSTM - Bi - 0,1 tfr)	5,897 (GRU - Mono - 0,1 tfr)
Perplexity	357,803 (LSTM - Bi - 0,1 tfr)	363,89 (GRU - Mono - 0,1 tfr)
BLEU-2	0,0057 (GRU - Bi - 0,5 tfr)	0,0086 (LSTM - Mono - 0,1 tfr)
BLEU-4	9,55e-81 (GRU - Bi - 0,1 tfr)	0,000154 (GRU - Mono - 0,5 tfr)

Figure 3: Sequence to Sequence w/ and w/out attention results.

4 Text-to-Text Transfer Transformer language model pre-trained on Italian Corpus

Motivated by the poor results obtained with the previous recurrent sequence to sequence model, the second milestone was to rely on a large scale pre-trained language model.

Specifically, after a preliminar phase of analysis of the current state-of-the-art pre-trained language models for text-generation, we decided to employ iT5, developed by Sarti and Nissim (2022), a huge language model pre-trained on a large italian corpus (275GB and 1 epoch of masked language modeling pre-traininig - approximately 1 mln steps), whose architecture and pre-training strategies are based on the famous Text-to-Text Transfer Transformer language model (Raffel et al. (2019)).

The model is an encoder-decoder transformer architecture developed by Vaswani et al. (2017) in 2017. The peculiarities of this architecture (Fig. 4) are:

- the use of a positional encoding along with the embedding encoding to augment the information provided to the model;
- the heavy use of the self-attention mechanism in order to extract information regarding the affection between all the words in the input;

- the use of masked attention in the decoder in order to generate at each step an output word based only on the previously generated ones and the encoded input (autoregressive generation).

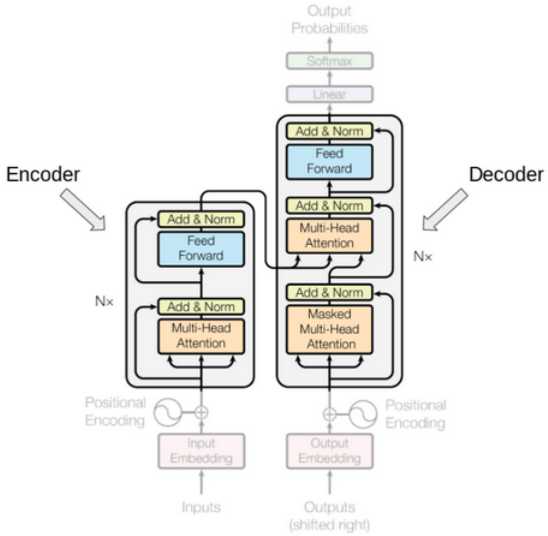
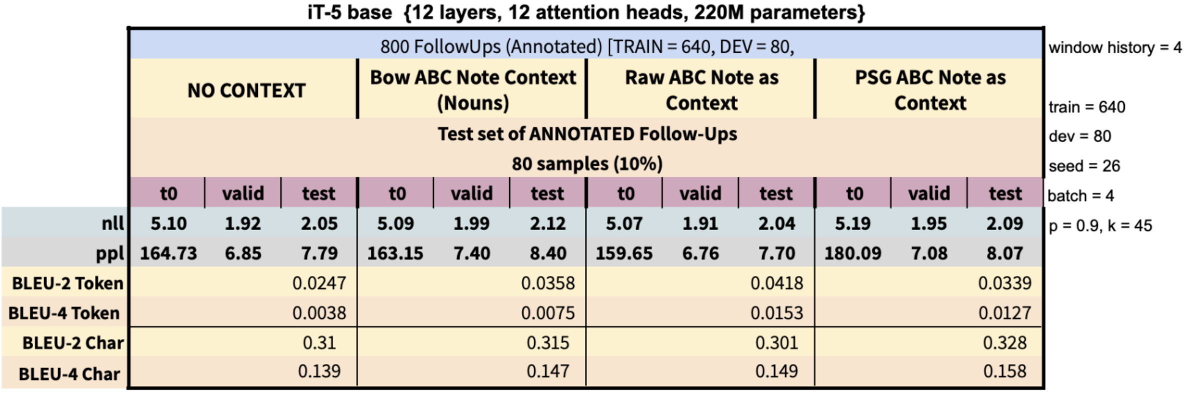


Figure 4: Transformer Architecture.

I fine-tuned and tested the three different version (small, base and large) of the iT5 architecture in order to find the best trade-off between the size of the model and the performance obtained. All the the models were trained using the AdaFactor optimizer (Vaswani et al. (2017), batch size equal to 4 (except for large one, whose batch size used was 2), dialogue history window of size 4 and early stopping.

Since the model, during the training phase, relies only on greedy decoding to generate text, during test phase we had to investigate the best set parameters for the decoding process (also called lexicalization), from the output distribution to the lexicalized output. Computing a grid search-like algorithm, I found the best decoding parameters to be $top - p = 0.9$ and $top - k = 45$.

Motivated by the results obtained (Fig. 5) and also by our computational availability, we decided to keep only the base version for the human evaluations steps.



iT-5 base {12 layers, 12 attention heads, 220M parameters}

	800 FollowUps (Annotated) [TRAIN = 640, DEV = 80,												window history = 4
	NO CONTEXT			Bow ABC Note Context (Nouns)			Raw ABC Note as Context			PSG ABC Note as Context			train = 640
	Test set of ANNOTATED Follow-Ups												dev = 80
	80 samples (10%)												seed = 26
	t0	valid	test	t0	valid	test	t0	valid	test	t0	valid	test	batch = 4
nll	5.10	1.92	2.05	5.09	1.99	2.12	5.07	1.91	2.04	5.19	1.95	2.09	p = 0.9, k = 45
ppl	164.73	6.85	7.79	163.15	7.40	8.40	159.65	6.76	7.70	180.09	7.08	8.07	
BLEU-2 Token			0.0247			0.0358			0.0418			0.0339	
BLEU-4 Token			0.0038			0.0075			0.0153			0.0127	
BLEU-2 Char			0.31			0.315			0.301			0.328	
BLEU-4 Char			0.139			0.147			0.149			0.158	

Figure 5: Results of the iT5 base model.

After this step, we packed up a file containing all the test dialogue histories along with the groundtruth and the generated responses from the iT5-Base model and from the GePpEtTo model Vaswani et al. (2017) (a language model based on GPT-2 architecture, developed by Radford et al. (2019), which was pre-trained on another Italian Corpus) with the four configurations of knowledge (None, Raw, BoW and PSG). With this final 9 responses per dialogue history we decide to automatically compare them using BLEU and Tf-IDF. For the first metric, we compute the average BLEU-2 and BLEU-4 scores between all the responses (Fig. 6) in order to obtain a first indication of the overlapping between the correct responses and the generated ones.

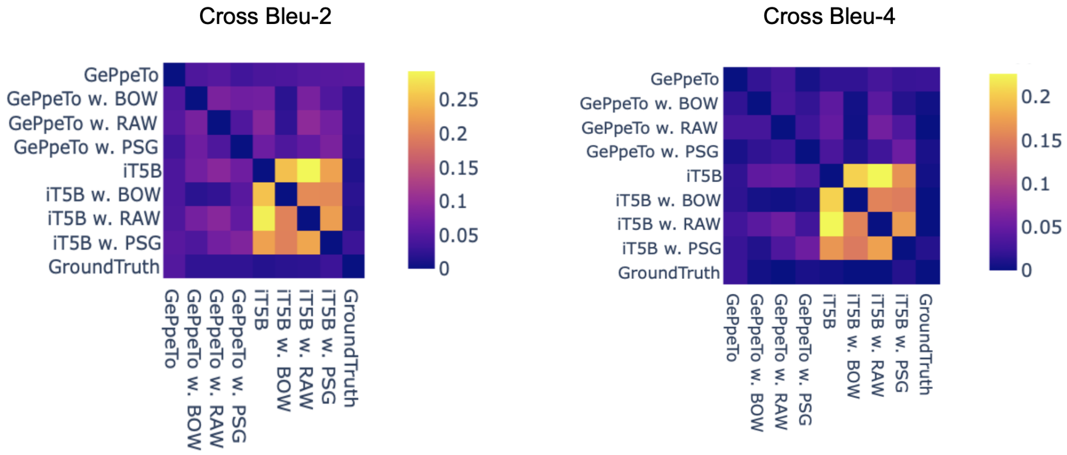


Figure 6: Cross BLEU-2 and BLEU-4 between all the responses.

Regarding the Tf-IDF, we employed it in order to gain a shallow feature level assessment of contextualization level of the generated responses. To do this, I trained a Tf-IDF model on the train-set and I used it to encode the dialogue histories, the groundtruth and the generated responses of the previously packed up file. Then, I computed the average cosine similarity between the dialogue history encoded vector and all the responses, obtaining the results in the plot 7.

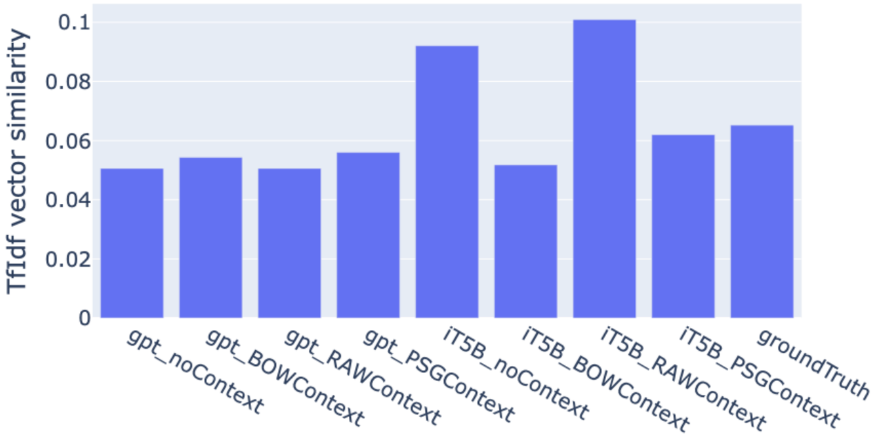


Figure 7: Tf-IDF cosine similarity between the encoded dialogue history and each response.

As a final evaluation, I repeatedly run the training and recorded the final test perplexity with incremental portions of the train-set (1%, 5%, 25%, 50%, 75% and 100%) extracted using stratified sampling over the different dialogues based on the dialogue history size. This was done to ensure a balanced subsampled dataset w.r.t. the original train-set. In the plot 8, it can be seen how the increase of the training size affects the performances, despite, at a certain point, it can be seen how the perplexity reaches a plateau where no improvements (or at least negligible ones) are recorded.

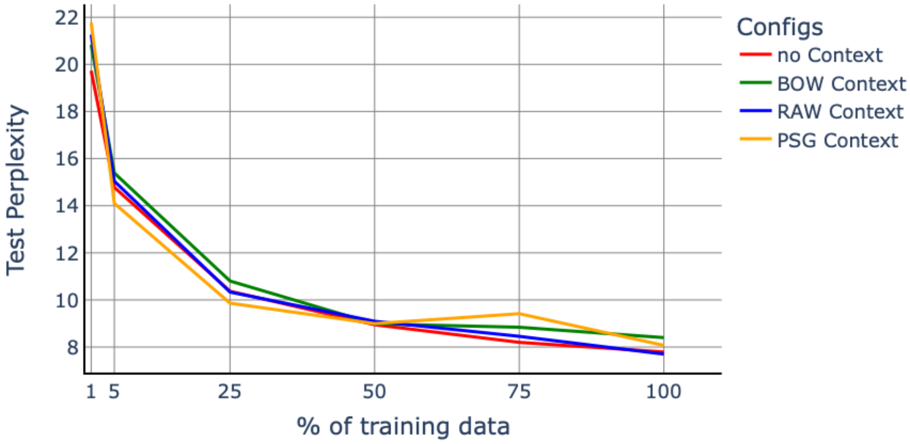


Figure 8: Text Perplexity obtained on model trained with incremental portion of the train set.

5 Conclusion

During this three month internship, co-supervised by (S.) Mahed Mousavi, I had the opportunity to deepen my knowledge in the field of Natural Language Understanding, and more specifically Dialogue Generation, developing and fine-tuning state-of-the-art language models. First, I implemented from scratch a complete Seq2Seq model with and without attention and assessed its performance. Second I fine-tuning the just-developed iT5 model and compared its performance with the previously developed GePpeTto model. In doing this, I also learnt about the common pitfalls and mistakes that occur during training and evaluation of deep learning models. All the work done is a part of an ongoing research entitled "Grounded Response Generation in the mental Health Domain".

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Mousavi, M. S., Negro, R., and Riccardi, G. (2022). An unsupervised approach to extract life-events from personal narratives in the mental health domain.

Mousavi, S. M., Cervone, A., Danieli, M., and Riccardi, G. (2021). Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9, Online. Association for Computational Linguistics.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.

Sarti, G. and Nissim, M. (2022). It5: Large-scale text-to-text pretraining for italian language understanding and generation.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need.