Improvement of Attention based Formality Style Transfer model with In-domain Data Augmentation

Korea University COSE461 Final Project

Jisoo Lee, Dayeon Ki
Department of Computer Science, Department of Statistics
Team 44
2019320062, 2019150419

Abstract

The inadequacy of training data is the most significant impediment to progress in the job of formality style transfer. We investigate how to augment parallel data and present several in-domain data augmentation strategies: synonym replacement and round trip translation. We approach the problem utilizing the attention based sequence to sequence model with three different scoring functions: dot, general and concatenate. We have also discovered quantitative evaluation metrics and human qualitative evaluations. Experiments with GYAFC benchmark dataset demonstrate that our augmented parallel data largely helps improve formality style transfer.

1 Introduction

Linguistic researchers have been intrigued by the field of text style transfer for a long time. Within the field, the interest for formality text style transfer(FST) has increased profoundly (Wang et al., 2019). Formality text transfer is to convert a formal piece of text into an informal piece of text, vice versa. Despite the wide interest in FST however, there still exists a lack of parallel corpora that could lead to good task performance. Recently, such a challenge has been tackled by Rao and Tetreault who created GYAFC corpus to strengthen the baseline model of FST. Inspired from the study, we aim to further investigate by conducting data augmentation methods that expect performance enhancement of FST model.

In this study, we take account of two data augmentation methods. First is synonym replacement and the other is round trip translation. Both of these methods focuses on in-domain training that avoids losing generality and assures the quality of data (Sosuke, 2018). Augmentation that replies on synonym replacement replaces certain words of the sentence and round trip translation translate the sentence from one language to another and back.

Machine transformers are often modeled through sequence to sequence(Seq2Seq) neural architecture. The language model leverages the likelihood of belonging to the target domain and predicts the next word (Etinger and Black, 2019). Nevertheless, one of the core issue of Seq2Seq model is the bottleneck problem, which refers to difficult situation for the model to capture the information of input vector as the length of sequence increases. Hence, we investigate the effects of data augmentations with Seq2Seq model along with the attention mechanism. We explore three different scoring functions that are dot, general and concatenate and evaluate our augmented model compared to the baseline model using BLEU score as our metric.

2 Related Work

Data augmentation has been a great ongoing topic of research for Seq2Seq tasks like Machine Translation (He et al., 2016; Edunov et al., 2018) and Grammatical Error Correction (Zhao et al., 2019; Zhou et al., 2019). For text style transfer, however, due to the lack of parallel data, many studies focus on unsupervised approaches (Luo et al., 2019; Wu et al., 2019) and there is little related work concerning data augmentation. As a result, most recent work that models text style transfer as Machine Translation suffers from a lack of parallel data for training, which seriously limits the performance of the proposed model.

Zhang et al. (2020) presented novel data augmentation methods for formality style transfer based on the Seq2Seq model architecture. The first method is back translation in French to obtain synthetic parallel sentences. Other methods also include formality discrimination, which generates formal rewrites of informal source sentence using cross-lingual MT models, and multi-task transfer that use annotated sentence pairs from other Seq2Seq tasks. The presented methods lack the focus on in-domain data augmentation although the importance of in-domain training is emphasized (Etinger and Black, 2019) in many NLP tasks. Hence, our work employed other data augmentation methods such as synonym replacement that consider in-domain training.

To solve the pain point of inadequacy of training data, we propose in-domain data augmentation methods and study the optimal way to utilize the augmented data, which not only achieves success in formality style transfer but also would be inspiring for other text style transfer tasks.

3 Approach

In this project, we attempt to train sequence to sequence(Seq2Seq) model with the attention mechanism. Seq2Seq is widely used for Machine Translation; however, it also plays a powerful role in the field of machine transformers. Considering the size of our sentence pairs which are long, we decided to apply the attention mechanism that gives greater weights to different parts of the input at distinct steps. Overall, we can break down our approach into 3 main sections: Data pre-processing, encoder-decoder framework with attention mechanism, and model pipelines.

3.1 Data Pre-processing

In the pre-processing level we adapt the method of using '<' and '>' symbols that indicates start and end of a text sequence, respectively. The input text of the encoder is closed under both symbols while the decoder input utilizes only the initiation symbol and the decoder output utilizes the termination symbol. Moreover, we tokenize both formal and informal data by allocating unique identifiers for each vocabulary within the given sentence. In the process we also exclude common punctuation such as ;?: etc. Lastly to adjust the lengths of the texts we carry out data padding that appends zeros to the sequences and randomize the input order. The max length sentence is explored through checking the distribution of input lengths. In our case we filter out sentences greater than 150.

encoder input	**decoder input**	**decoder output**
<I do not intend to be mean.>	<I don't want to be mean.	I don't want to be mean.>
<I mean that you have to really be her friend.>	<and i mean Really be her friend.	and i mean Really be her friend.>

Table 1: Examples of tokenized data

3.2 Encoder-Decoder Framework with Attention Mechanism

The encoder of our neural architecture is comprised of three Unidirectional LSTM layers. The encoder takes two arguments as a tuple. The first index refers to the sequence input and the other to

the initial states of the encoder. The encoder returns the recent step's hidden state and current state. Refer to Figure 2 for the diagram.

The crucial role of our exploration takes part in the attention computation. Different scoring functions results in different weights to the output encoder tokens. The three scoring functions of attention mechanism is applied depending on which string of argument has been passed to the attention class. For instance, "dot" indicates to implement dot-product attention.

$$f(h_{t-1}, e_{t'}) = \begin{cases} h_{t-1}^T e_{t'} & \text{dot} \\ h_{t-1}^T W e_{t'} & \text{general} \\ v^T \tanh(W[h_{t-1}, e_{t'}]) & \text{concat} \end{cases}$$

Figure 1: Three scoring functions of the attention mechanism : dot, general and concatenate

The attention decoder provokes the time-step decoder before producing its output. The time-step decoder holds two Unidirectional LSTM Decoder and a Dense layer. The Dense layer operates to create fully connected layers that weight the decoder output according to the size of the output vocabulary. Correspondingly, the time-step decoder combines the output of the preceding decoder time-step with the attention weights generated by the attention model.

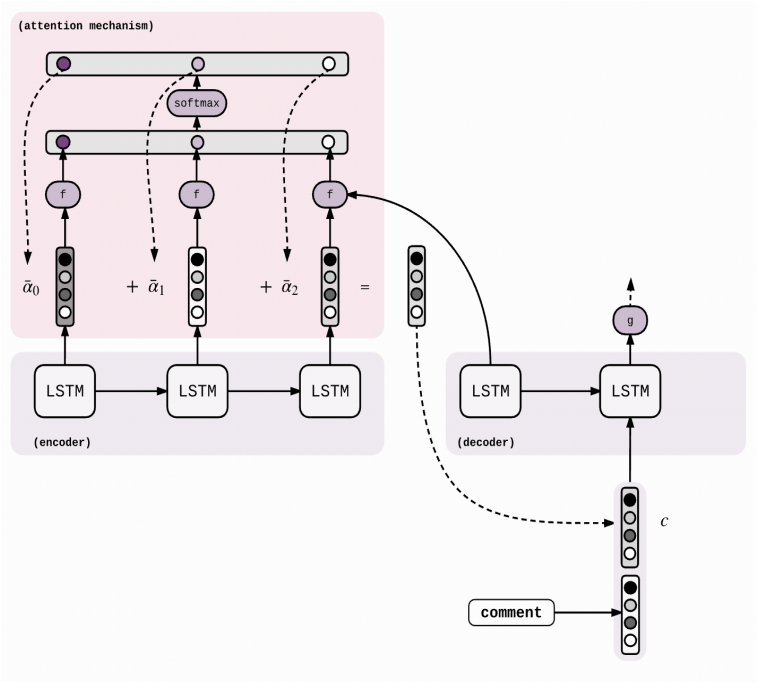

Figure 2: Diagram of Seq2Seq model with attention mechanism

3.3 Model Pipeline

To ensure better performance of our model training, we consider several functions. First, we construct the loss function and customize it so that it masks the zero-padded parts. Secondly, the callback function provides a log directory to keep track of the loss. Lastly, we utilize model checkpoints that stores the best weights. The checkpoints were important especially because we work with Google Colab that puts limits to the GPU usage and were required to restore the trained model.

4 Experiments

4.1 Data

We use the baseline dataset as the Grammarly's Yahoo Answers Formality Corpus(GYAFC) provided by Rao and Tetreault (Rao and Tetreault, 2018). Additionally, we create two datasets that are also based from GYAFC corpus but have been augmented through synonym replacement or round trip translation. To get admission to the corpus, we first requested access to the Yahoo Answers corpus and then we forwarded the acknowledgment to one of the authors.

GYAFC Dataset

GYAFC is one of the largest parallel datasets of human-labeled style transfer between formal and informal sentences. Rao and Tetreault, the authors of the dataset, extracted informal sentences from Yahoo Answers L6 corpus to create the sentence pairs. The Yahoo Answers L6 corpus is comprised of several different domains but GYAFC considers only two that possess a greater amount of informal sentences which are *Entertainment & Music (E&M)* and *Family & Relationships(F&R)*. The distribution of the data has been carefully considered at the preprocessing step such as removing questions, URLS, or words shorter than 5 and longer than 25 etc. Formal sentences were produced with mechanical turks and have been reviewed by experts repetitively. Overall, GYAFC contains training sentence pairs of 50k, validation of 3k, and testing of 1.5k for both domains.

		Informal to Formal		**Formal to Informal**	
	Train	Tune	Test	Tune	Test
E&M	52,595	2,877	1,416	2,356	1,082
F&M	51,967	2,788	1,332	2,246	1,019

Table 2: GYAFC dataset statistics

Informal: *I'd say it is punk though.* Formal: *However, I do believe it to be punk.*
Informal: *Gotta see both sides of the story.* Formal: *You have to consider both sides of the story.*

Table 3: Sentence examples of GYAFC dataset

Synonym Replacement Dataset

We utilize synonym replacement as one of our data augmentation methods. The augmentation is implemented through replacing a certain number of words in the formal sentence with their synonyms through WordNet. The replacement considers the cosine similarity and computes the likeness. Overall, synonym replacement results in about 210k training dataset appended with the baseline dataset. Examples of augmentation are shown below.

Original: *I have never seen the show but I am worried that she will win because she is a Scientologist* Augmented: *I have never take in the show but I am distressed that she will pull ahead because she is a Scientologist*
Original: *i can't sign in at the chatroom* Augmented: *I am unable to sign in at the chat room.*

Table 4: Sentence examples of synonym replacement augmented GYAFC dataset

Round-trip translation Dataset

Another technique applied to the baseline dataset is round trip translation. We translate from English to the target language and translate back to English. The target language used in our paper is French that is one of the widely used languages for round trip translation. The resultant of the RTT-based metric is 210k training dataset.

Original: *Perhaps it is because men are afraid to hurt girl's feelings.* Augmented: *Maybe because men are afraid of hurting girls' feelings.*
Original: *I presume it to be so, but avatars make anyone look great.* Augmented: *I guess that's the case, but the avatars make everyone look good.*

Table 5: Sentence examples of round-trip translation augmented GYAFC dataset

4.2 Evaluation method

For quantitative analysis, we implemented BLEU (Bilingual Evaluation Understudy) metric for evaluating the generated informal sentence. BLEU score quantifies the quality of generated sentences from one natural language to another by calculating the correspondence between the model's predicted output and the gold answer. We use the score that is averaged over the whole corpus of generated sentences to reach an estimate of the model's overall quality. Based on the BLEU score of each generated sentence of the model, we examine the top 5 best and worst predictions to pursue error analysis.

4.3 Experimental details

In this section, we present the experimental settings and related experimental results. We focus on formal to informal style transfer since it is more practical in real application scenarios.

For the experimental settings, we use the attention-based Transformer (Vaswani et al., 2017) as the Seq2Seq model with a shared vocabulary of 20K BPE (Sennrich et al., 2016b) tokens. After carefully examining the distributions of lengths of encoder input, decoder input, and decoder output, we realize that almost all the sentences have lengths less than 150. Hence, we filter out the sentences which are of length more than 150. We adopt Adam optimizer and the learning rate is initially set as 0.01. Throughout the training, when the metric stops improving, the learning rate is reduced with a minimum value set as 0.0001. We train the model for a total of 12 epochs with the option of early stopping.

4.4 Results

Model	**Scoring Function**	**BLEU Score**
Baseline	Dot General Concat	0.308 0.269 0.202
Synonym Replacement (SR)	Dot General Concat	0.313 0.350 0.313
Round-trip Translation (RTT)	Dot General Concat	**0.358** 0.332 0.300

Table 6: Final results (12 epochs)

The result presents that data augmentation overall improves the model performance of text formality transfer. Comparing the BLEU scores, the baseline has relatively lower scores than other augmented

models that comes up to our expectations. Data augmentation methods make use of greater diverse examples of training datasets leading the model to learn more rich and sufficient expressions of the input sequence.

5 Analysis

We compare the BLEU scores of three scoring functions: dot, general, and concatenate. Results in Table 6 suggest that for the dot scoring function, the BLEU score of the baseline is 0.308, 0.313 for the SR, and 0.358 for RTT with the highest value. For the general scoring function, the baseline BLEU score is 0.269, SR is 0.350 and RTT is 0.332. For concatenate scoring function, the BLEU score for the baseline is 0.202, 0.313 for SR, and 0.300 for RTT. BLEU score increased by applying data augmentation for all three scoring variants.

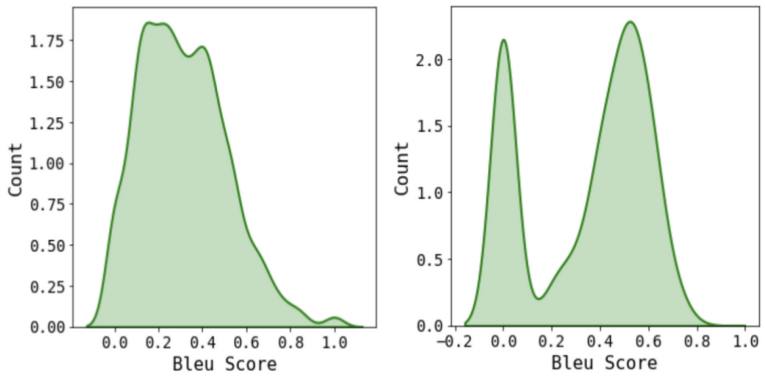

Figure 3: BLEU score distribution with averaged score of 0.313 (left, SR with dot scoring function) and 0.350 (right, SR with general scoring function)

We realize that solely observing the BLEU score was insufficient for explaining models with skewed BLEU score distribution as in Figure 3. Although the model on the left has a higher BLEU score than the right, the overall distribution is skewed. This discrepancy occurs as we simply use the averaged score of the whole corpus.

Therefore, we also consider the loss calculated through our custom loss function. Our loss function is based on categorical cross entropy but modified to not consider the losses of the padded zeros. The training and validation loss values not only verify the BLEU scores but also show whether the model is overfitting.

Model	**Scoring Function**	**BLEU Score**	**Train loss**	**Validation loss**
Baseline	Dot General Concat	0.308 0.269 0.202	0.302 0.519 0.498	0.400 0.490 0.514
Synonym Replacement (SR)	Dot General Concat	0.313 0.350 0.313	0.347 0.419 0.347	0.369 0.452 0.369
Round-trip Translation (RTT)	Dot General Concat	0.358 0.332 0.300	0.348 0.459 0.565	0.351 0.449 0.546

Table 7: Training and Validation loss values (12 epochs)

To overcome the limitations of quantitative evaluation, we also analyze qualitative aspects of the results. For the highest scoring variant of each model (baseline, SR, RTT), we conduct error analysis by examining the top 5 best and worst predictions. Below is an example of the generated predictions and the gold answer.

Best Prediction: *i traveled there and he was present.* Expected Output: *i went and there he was.*
Worst Prediction: *cause it's buy one take one.* Expected Output: *its because is because it is belle it is belle it is belle it is belle*

Table 8: Error Analysis of Best and Worst predictions

There is a general pattern for the worst predictions such as constant repetition of a certain word/ phrase or generating an informal sentence with a completely different context. Interesting examples of error analysis also include cases when the formal input gives a sentence about a 'girl' when the informal output gives a sentence about a 'boy'. This shows that in some cases, the model is not fully understanding the input context and giving random predictions.

6 Ablation Study

6.1 Analysis of Epoch

Observations on the BLEU score distributions have shown us that the BLEU score itself is an insufficient method to evaluate model performance. If the distribution is skewed, it is unclear whether we can simply justify the performance by the averaged BLEU score. To solve this problem, we experimented with models of similar BLEU scores but different score distributions by changing the number of training epochs. These models include SR augmented model (dot, general scoring function) and RTT augmented model (dot, general scoring function). We observe the changing BLEU scores as the number of epochs increase from 5 to 20.

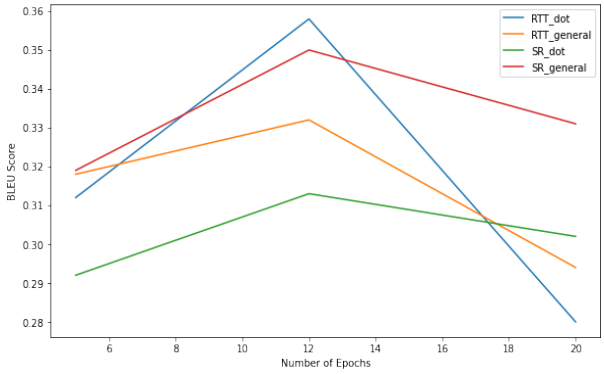

Figure 4: Change in BLEU scores by different number of epochs

As shown in Figure 4, BLEU scores decrease beyond epoch 12 for all cases. The cause for the decrease in the BLEU score was the overfitting problem. This means that the number of epochs used to train was more than necessary, leading to model learning patterns specific to train data. The training loss was smaller than validation loss to a great extent confirms the overfitting.

6.2 Analysis of Pivot Languages

Besides modifying the number of epochs, we also explored the effect of pivot languages for round trip translation. Three pivot languages we focus on are French, Spanish and Chinese.

Model	Scoring Function	French (Fr)	Chinese(Zh)	Spanish(Es)
BLEU Score	Dot	0.358	0.316	0.2719
	General	0.332	0.300	0.301
	Concat	0.300	0.304	0.302

Table 9: Performances of round trip translation on different languages: French (Fr), Chinese (zh) and Spanish (Es)

In general, based on the observations from 12 epochs of training, the French outperformed. Based on human error analysis it is inspected that French presented better results due to practically identical sentence structure of English and French (Zhang et al, 2020). On the other hand, Chinese and English has quite different structures that create more noise during the translation. Eventually this may have altered the meaning and the formality. For instance, Spanish grammatically differ from English heavily in terms of infection and declination (Ahmadnia et al., 2017).

7 Conclusion

In this paper, we propose in-domain data augmentation methods for formality style transfer. Our proposed data augmentation methods can effectively generate diverse augmented data with various formality text style transfer knowledge. The augmented data can significantly help improve the performance of the proposed attention-based Seq2Seq model. We also compare the effect of attention variants including dot, general, and concatenate scoring functions.

For further investigation, we suggest limitations and possible improvements to our work. First, we only consider one-way generation from formal to informal. It would be more meaningful to extend our research by considering bidirectional transformation between informal and formal. Regarding each aspect of the model structure, we could improve the encoder by applying Bidirectional LSTM rather than Unidirectional LSTM. It allows the model to take advantage of the bidirectional context of the input sequence. We could also improve the decoder part by adding stopping criteria such as beam search decoding. It would extract the context or phrase similarities to greater extent. Furthermore, although our work only considered the Seq2Seq model with attention mechanism, we hope that using pre-trained language models such as BERT (Devlin et al., 2018) or GPT (Brown et al., 2020) can bring substantial improvements of the performance.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ahmadnia, Benyamin, Javier Serrano, en Gholamreza Haffari. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. Varna, Bulgaria: INCOMA Ltd., 2017. 24–30.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers,* pages 4873–4883. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems,* pages 820–828.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019,* pages 5116–5122. ijcai.org.

Genthial, Guillaume. "Seq2Seq with Attention and Beam Search." Guillaume Genthial Blog, 8 Nov. 2017, guillaumegenthial.github.io/sequence to sequence.html.

Isak Czeresnia Etinger, and Alan W. Black. 2019. Formality Style Transfer for Noisy Text: Leveraging Out-of-Domain Parallel Data for In-Domain Training via POS Masking. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), Hong Kong, China,* pages 11–16. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanove. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Kobayashi, Sosuke. 2018. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. *arXiv preprint arXiv:1805.06201*

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.*

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back translation at scale. *arXiv preprint arXiv:1808.09381.*

Sudha Rao and Joel R. Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers),* pages 129–140.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165.*

Wangchunshu Zhou, Tao Ge, Chang Mu, Ke Xu, Furu Wei, and Ming Zhou. 2019. Improving grammatical error correction with machine translation pairs. *arXiv preprint arXiv:1911.02825.*

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138.*

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel Data Augmentation for Formality Style Transfer. *arXiv preprint arXiv:2005.07522.*

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3573–3578.

A Appendix: Team contributions

Jisoo Lee formulated preprocessing and encoder code, progressed synonym replacement experiment and RTT Spanish experiment for ablation study. Dayeon Ki formulated decoder and training code, conducted baseline experiment and RTT Chinese experiment for ablation study. Cooperated with each other for related work research, RTT French experiment and preparation for the final report.