THE TREATMENT OF ERRORS MADE BY FRENCH SECOND LANGUAGE LEARNERS IN THE USE OF OBJECT CLITIC PRONOUNS THROUGH THE USE OF A FINE-TUNED DEEP LEARNING MODEL

Adel Jebali

Concordia University, Montreal, Canada

Email: adel.jebali@concordia.ca

BIOGRAPHY

Adel Jebali is an associate professor in the Department of French Studies at Concordia University in Montreal, Quebec, Canada. With a PhD in linguistics specialized in computational linguistics, his research projects focus on natural language processing and its integration into second language teaching. The linguistic phenomenon he specializes in is that of French clitic pronouns.

OUTLINE

- Introduction / background
- CamemBERT
- Corpus
- Fine-tuned model
- Conclusion

INTRODUCTION / BACKGROUND

- Object clitic pronouns in French (OCs):
  - Example: Marie a mangé la pomme. → Marie **l'**a mangée.
    - In the example: **l'** is an OC
    - This OC has to agree with its antecedent: gender, number, person and animacy features.
    - This OC: placed before the verb (and the auxiliary) while the corresponding NP is placed postverbally.
  - Difficulties in mastering for second language learners : strategies include avoidance (NP is repeated, for example)
  - In the French L2 writings: a small amount of Ocs are observed.
  - Elicitation tasks allow researchers to have more production (Jebali, 2018).
  - My goal: help French L2 learners master these OCs by automatically correcting their production.

INTRODUCTION / BACKGROUND

- Current French grammar checkers: unsatisfactory regarding OCs
- Antidote, for example: context is not taken into account
- The context: crucial for the mastery of OCs (agreement with the antecedent)
- My goal: to develop another tool aiming OCs in particular
- Deep learning models: proved to be efficient in natural language tasks, such as classification, parts of speech tagging, named entities recognition, etc.
- I chose a state-of-the-art deep learning model to fine-tune on my data: CamemBERT

CAMEMBERT

- Transformer-based (Vaswani et al, 2017)
- CamemBERT, described in (Martin et al 2020)
- A state-of-the-art deep-learning language model for French based on Bidirectional Encoder Representations from Transformers (BERT) and more specifically on RoBERTa
- 110 million parameters
- Was pretrained on the French subcorpus of the multilingual corpus OSCAR (138 GB of text) as part of a collaboration between INRIA Paris (ALMANACH team) and Facebook/Meta AI
- My contribution: fine-tune CamemBERT on my corpus (Jebali, 2018).

CORPUS

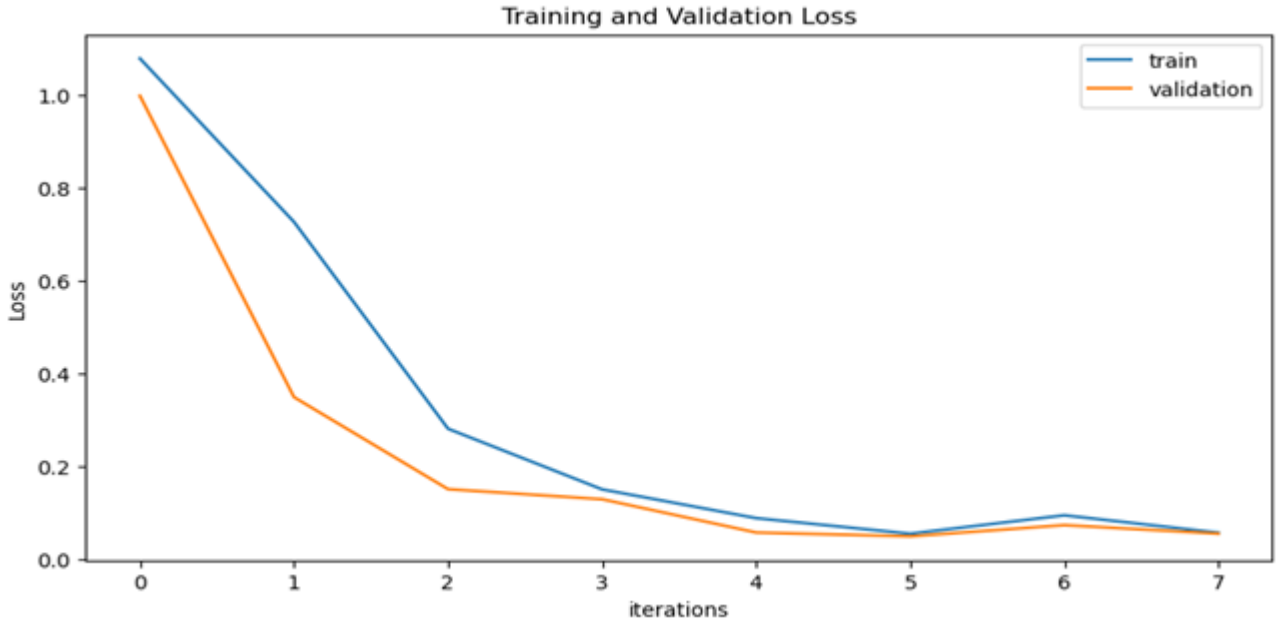
- The dataset used to fine-tune the model: from a previous research project on new technologies and their quantitative and qualitative effects on the production of French L2 OCs
- The corpus: as a basis to isolate both the OC and a relevant context of its use
- The interview-like nature of the corpus → pairs containing a question and the answer to it. Examples:
  1) Qu'est-ce que j'ai fait avec mes crayons? Tu les as rangés. (English: What did I do with my pencils? You put them away.)
  2) Que fait la fille avec cette pomme? La fille épluche la pomme. (English: What is the girl doing with this apple? The girl is peeling the apple.)
  3) Qu'est-ce que j'ai fait avec mon crayon? *Tu as l'aiguisé. (English: What did I do with my pencil? *You have sharpened it.)
- (1): correct, (2): repeated NP, (3): error in the OC placement

CORPUS

- Dataset to fine-tune the model: 899 pairs
  - 437 pairs with repeated NPs
  - 336 pairs with correct answers
  - 126 pairs with errors
- Unbalanced dataset: the Weighted Random Sampler from the Pytorch library used to give the less represented data a weight based on their size.
- 80%: training, 20%: validation

FINE-TUNED MODEL

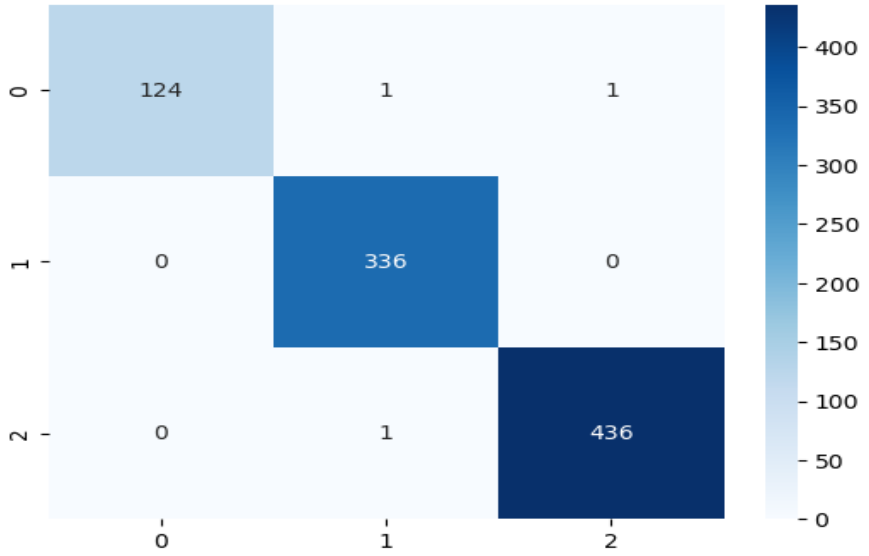
- I trained CamemBERT for 10 epochs on an Nvidia GPU with AdamW as the optimizer
- The early stop technique was used: stopped the training at epoch 7



FINE-TUNED MODEL: CLASSIFICATION REPORT / CONFUSION MATRIX

- Overall f-score: 0.99

	Precision	Recall	F1-score	Support
Error	1.00	0.98	0.99	126
Correct	0.99	1.00	1.00	336
Repetition	1.00	1.00	1.00	437



FINE-TUNED MODEL

- The model: performed very well on the validation data, with only three pairs misclassified
- The model: was able to predict the appropriate grammatical judgment for pairs that it had never seen before but appeared similar to the training data
  - Example: Qu'est-ce que j'ai fait avec la carte? Tu as sorti la carte.
  - English: What did I do with the card? You took the card out.
  - Correct prediction: 2 (repetition)
- It was also able to correctly make predictions in different contexts of OCs usage
  - Example: *Je lui aide.
  - English: I help him/her.
  - Correct prediction: 0 (error)

CONCLUSION

- Development of a graphical interface in Python and PyQT6 to interact with this model
- This interface: only provides a label and a recommendation
- In a future version: improve the predictive capabilities of the model as well as the feedback refined by error type.
- Improve predictive capabilities:
  - More errors in the dataset, with different types
  - A balanced dataset
  - A refined feedback about errors