Evaluating Deep Learning Methods for Tokenization of Space-less texts in Old French and Latin

Thibault Clérice[1]

[1]École nationale des Chartes, France
[2]Université Lyon 3, France

Corresponding author: Thibault Clérice , thibault.clerice@chartes.psl.eu

Abstract

Tokenization of modern and old Western European languages seems to be fairly simple as it stands on the presence mostly of markers such as spaces and punctuation. Although, when dealing with old sources like manuscript written in *scripta continua*, (1) such markers are mostly absent, (2) spelling variation and rich morphology makes dictionary based approaches difficult. We show that applying convolutional encoding to characters followed by linear categorization to word-boundary or in-word-sequence can be used to tokenize such inputs. Additionally, we release a software with a rather simple interface for tokenizing one's corpus.

Keywords

convolutional network; scripta continua; tokenization; Latin; Old French; word segmentation

I INTRODUCTION

Tokenization of space-less strings is a task that is specifically difficult for computer when compared to "whathumancando". *Scripta continua* is a writing phenomenon where words would not be separated by spaces and it appears to have disappeared around the 8th century (see Zanna [1998]). Although, spacing can be somewhat erratic in later centuries (*cf.* Figure 1). In the context of text mining of HTR or OCR output, lemmatization and tokenization of medieval western languages can be a pre-processing step for further research to sustain analyses such as authorship attribution **CITE JBCAMPS ?**.

We must stress in this study that the difficulty that we face is different for *scripta continua* than

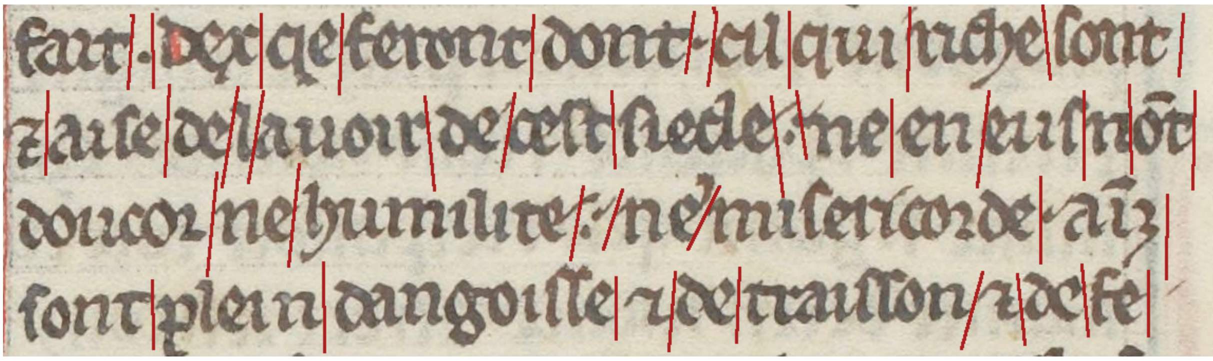
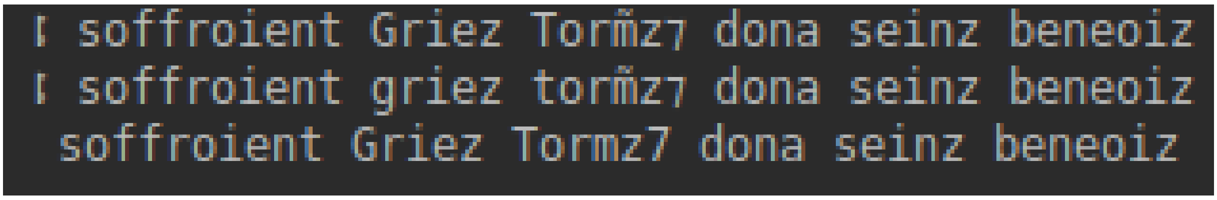


Figure 1: 4 lines from fol.103rb Manuscript fr. 412, Bibliothèque nationale de France. Red lines indicate word boundaries

	Sample
Input String	`Ladamehaitees’enparti`
Mask String	`xSxxxSxxxxxSxxxSxxxxS`
Output String	`La dame haitee s’en parti`

Table 1: Input, mask and human-readable output generated by the model. x are WC and S are WB



```
 soffroient Griez Torm̃z⁊ dona seinz beneoiz
 soffroient griez torm̃z⁊ dona seinz beneoiz
 soffroient Griez Tormz7 dona seinz beneoiz
```

Figure 2: Different possibilities of pre-processing. In the last, the first character (LATIN ABBREVIATION SIGN SMALL ET WITH STROKE) is lost by `unidecode`

the ones researchers face languages such as Chinese for which an already impressive amount of work has been done as it. Indeed, Chinese word segmentation has lately been driven by deep learning methods, specifically ones based on *sequence to sequence translations*: Chen et al. [2015] defines a process based on LSTM model, while Yu et al. [2019] uses BiDirectional GRU and CRF. Actually, meanwhile redacting this article and producing the code-base, Huang et al. [2019] took the same approach of encoding to linear classification to both word boundary (WB) and word content (WC) for Chinese word segmentation.

II DESCRIPTION AND EVALUATION

2.1 Architecture

2.1.1 Encoding of input and decoding

The model is based on traditional text input encoding where each character is transcoded to an index. Output of the model is a mask that needs to be applied to the input: in the mask, characters are classified either as word boundary or word content.

For evaluation purposes, and to reduce the number of input classes, we propose both a lowercase normalization and a ”reduction to the ASCII character set” feature (fr. 2). On this point, a lot of issues were found with transliteration of MUFI characters that were part of the original datasets, as they are badly interpreted by the `unidecode` python package. Indeed, `unidecode` will simply remove characters it does not understand. An addendum to said package that would transcribe ligatures, abbreviations, etc. would probably be a good feature.

2.1.2 Model

Each model we have used is composed of an Encoder, generally built around an embedding layer, and the same Linear Classifier. The encoders tested are:

- LSTM encoder with hidden cell
- Convolutional (CNN) encoder with position embeddings
- Convolutional (CNN) encoder without position embeddings

2.2 Evaluation

2.2.1 *Datasets*

Datasets are transcription from manuscripts with unresolved abbreviation coming from different projects

- **Old French** based on Bluche et al. [2017], Pinche [2017], Jean-Baptiste-Camps et al. [2019], **?**, and transcription from the TNAH master at the École Nationale des Chartes.
  - 193,734 training examples;
  - 23,581 validation examples;
  - 25,512 test examples

2.2.2 *Results*

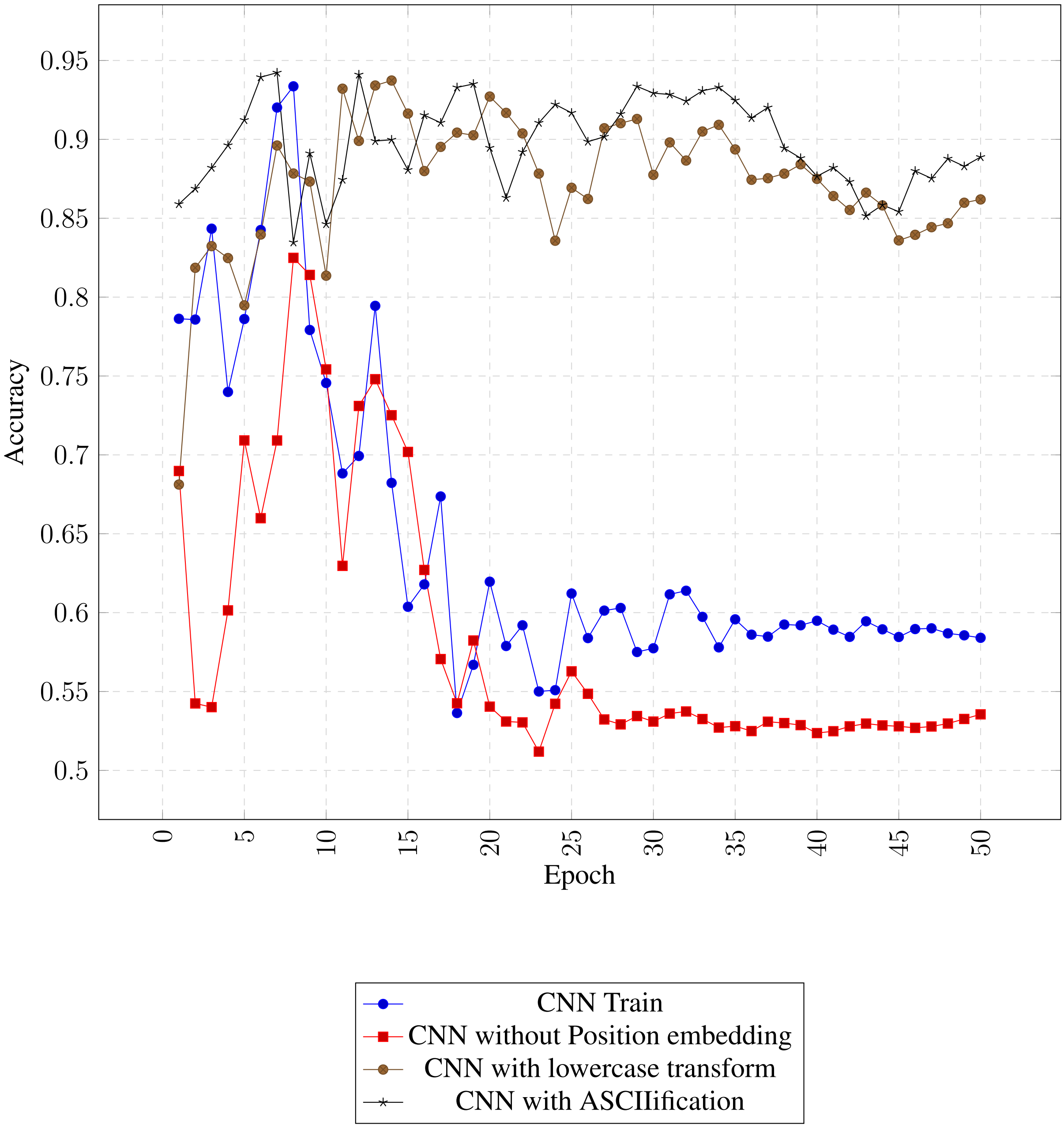


Figure 3: Training Accuracy over 50 epochs

2.2.3 *Example of outputs*

The following input was not seen during training, dev or test, and is part of a different corpus :

- `Truth` : Aies joie et leesce en ton cuer car tu auras une fille qui aura .i. fil qui sera de molt grant merite devant Dieu et de grant los entre les homes.
- `Input` : Aiesjoieetleesceentoncuercartuaurasunefillequiaura.i.filquiserademoltgrantmeritedevantDieu
- `CNN` : Aiesjoie et leesce en ton cuer car tu auras une fille qui aura . i . fil qui sera de molt grant merite devant Dieu et de grant los entre les homes .
- `CNN lower`: Aies joie et leesce en ton cuer car tu auras une fille qui aura . i . fil qui sera de molt grant merite devant Dieu et de grant los entre les homes .
- `CNN without position`: Aiesjoie et leesce en ton cuer car tu auras une fille qui aura . i . fil qui sera de molt grant merite devant Dieu et de grant los entre les homes .
- `CNN Normalize`: Aies joie et leesce en ton cuer car tu auras une fille qui aura . i . fil qui sera de molt grant merite devant Dieu et de grant los entre les homes .

2.3 Discussion

We believe that, aside from a graphical challenge, word segmentation in OCR from manuscripts can actually be treated from a text point of view

2.4 Conclusion

While

2.5 Acknowledgement

Boudams has been made possible by two open-source repositories from which I learned and copied bits of implementation of certain modules and without which none of this paper would have been possible: Manjavacas et al. [2019] and Trevett [2019]. This tool was originally intended for post-processing OCR for the presentation Camps et al. [2019] at DH2019 in Utrecht.

References

T. Bluche, S. Hamel, C. Kermorvant, J. Puigcerver, D. Stutzmann, A. H. Toselli, and E. Vidal. Preparatory kws experiments for large-scale indexing of a vast medieval manuscript collection in the himanis project. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 01, pages 311–316, Nov 2017. doi: 10.1109/ICDAR.2017.59.

Jean-Baptiste Camps, Thibault Clérice, and Ariane Pinche. Stylometry for noisy medieval data: Evaluating paul meyer's hagiographic hypothesis. In *DH2019*, July 2019.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1197–1206, 2015.

Chu-Ren Huang, Ting-Shuo Yo, Petr Simon, and Shu-Kai Hsieh. A realistic and robust model for chinese word segmentation, 2019.

Jean-Baptiste-Camps, LAKME-ENC, AliceCochet, LucenceIng, and Paulinelvq. Jean-Baptiste-Camps/Geste: Geste: un corpus de chansons de geste, 2016-..., April 2019. URL https://doi.org/10.5281/zenodo.2630574.

Enrique Manjavacas, Thibault Clérice, and Mike Kestemont. emanjavacas/pie v0.2.3, April 2019. URL https://doi.org/10.5281/zenodo.2654987.

Ariane Pinche. Édition nativement numérique des oeuvres hagiographiques 'Li Seint Confessor' de Wauchier de Denain, d'après le manuscrit 412 de la bibliothèque nationale de France. 40 ans du laboratoire du CIHAM et de la création du pôle de Lyon de l'EHESS, October 2017. URL https://hal.archives-ouvertes.fr/hal-01628533. Poster.

Ben Trevett. Pytorch seq2seq, April 2019. URL https://github.com/bentrevett/pytorch-seq2seq.

Chenghai Yu, Shupei Wang, and Jiajun Guo. Learning chinese word segmentation based on bidirectional gru-crf and cnn network model. *International Journal of Technology and Human Interaction (IJTHI)*, 15(3):47–62, 2019.

Paolo Zanna. Lecture, écriture et morphologie latines en irlande aux viiè et viiiè siècles. *Archivum Latinitatis Medii Aevi-Bulletin du Cange (ALMA)*, 1998.

A ANNEX 1

Pellentesque dignissim ultrices fringilla. Vivamus eu luctus ante, vel bibendum magna. Curabitur elit purus, tincidunt non dui vitae, elementum bibendum neque. Curabitur ullamcorper sit amet justo at hendrerit. Fusce ut arcu imperdiet nibh mollis tempus a aliquet tellus. Quisque pharetra cursus nisi, vel lobortis ante consectetur et. Vivamus sed congue neque. Proin pellentesque risus nec dui consequat rutrum. Vestibulum nunc diam, placerat quis auctor vel, faucibus non justo. Etiam dictum purus neque. Phasellus imperdiet mauris ligula, eu laoreet nisi elementum ut. Sed sed porta massa. Aenean faucibus risus ultrices ornare porta. Quisque faucibus ante a tincidunt vestibulum. Lorem ipsum dolor sit amet, consectetur adipiscing elit.

B ANNEX 2

Cras tristique vel nisi at aliquet. Proin egestas erat sit amet velit lobortis imperdiet. Integer et arcu sapien. Etiam id blandit sapien. Nam tempus lacus ac massa semper, vel laoreet turpis rutrum. Mauris eget nibh vitae justo porta imperdiet sed vel ligula. In imperdiet, augue vel condimentum convallis, neque augue imperdiet neque, eget dapibus nunc mauris ultricies tortor. Nam eget nunc egestas, blandit lectus non, aliquam nunc. Cras sed quam vitae arcu ornare lobortis. Ut ut lacus hendrerit, convallis orci sit amet, commodo nunc. Pellentesque eget tincidunt tortor. Nunc ornare molestie mauris id vehicula. Suspendisse pharetra tortor metus, sit amet fermentum tellus vehicula ut.