Universität Zürich[UZH]

TrOCR meets CharBERT

Masterarbeit der Wirtschaftswissenschaftliche Fakultät der Universität Zürich

eingereicht von

Yung-Hsin Chen

Matrikelnummer 20-744-322

Institut für Informatik der Universität Zürich

Prof. Dr. Martin Volk

Institut für Computerlinguistik der Universität Zürich

Supervisor: Dr. Simon Clematide, Dr. Phillip Ströbel

Abgabedatum: tbd

Abstract

This is the place to put the English version of the abstract.

Zusammenfassung

Und hier sollte die Zusammenfassung auf Deutsch erscheinen.

Acknowledgement

I want to thank X, Y and Z for their precious help. And many thanks to whoever for proofreading the present text.

Contents

List of Figures

List of Tables

List of Acronyms

OCR	Optical Character Recognition
LM	Language Model
NLM	Noisy Language Model

1 Introduction

1.1 Motivation

1.2 Research Questions

In this study, we aim to investigate and address the following key research questions.

1. In the Candidate Fusion paper, they claimed that fusing the recogniser and LM can make the LM adjust to the domain-specific data. Can fusing TrOCR and CharBERT achieve the same conclusion? In other words, can CharBERT adjust to historical texts even it was trained on modern texts?
2. In the CharBERT paper, they claimed that, by using the character level information in addition to the subword level information, the problems of incomplete modelling and fragile representation can be solved. However, the results shown in the paper did not show significant performance improvement over RoBERTa (a strong baseline LM model). Is this statement valid? Can TrOCR combined with CharBERT achieve the claim?
3. Which layers should be frozen to results in better performance and why?
4. Does TrOCR decoder change its beam search output with the presence of the language model?

2 Related Work

2.1 TrOCR

2.2 Candidate Fusion

3 Methodology

In this chapter, I will introduce the model which consists of TrOCR and CharBERT by combining the ideas of CharBERT and Candidate Fusion mentioned in chapter 2. First, I will describe the design concept of the model. Next, I will elaborate on the architectures of the models developed in this study.

3.1 Design Concept

In chapter 2, CharBERT and Candidate Fusion are introduced. CharBERT mitigates the problems of incomplete modelling and fragile representation by including the character encoding in addition to the subword level information. Furthermore, having NLM as the pre-training task makes CharBERT effective at correcting character level typos, which is a desired feature for post-OCR correcting.

Even though CharBERT outperforms RoBERTa most of the time, the improvements are not significant. RoBERTa is still considered a strong LM, comparing to CharBERT. Thus, it is still necessary to inspect the power of CharBERT by replacing it with RoBERTa and observe the results.

On the other hand, Candidate Fusion claims that having an interaction between the recogniser and the LM can enhance the performance of OCR. Thus, having TrOCR as the recogniser and CharBERT as the LM, combining them is expected have an improvement on the OCR accuracy.

CharBERT serves as a corrector in the model. However, it does not have strong semantic understandings. Thus, an additional LM with strong semantic understandings can also be added to the recogniser to perfect the model's capabilities in image recognition, semantics and robust in the presence of typos.

3.2 Model Architecture

To fully explore the effect of each component, namely RoBERTa, CharBERT and the language model, on the recogniser, TrOCR, an ablation study will be carried out. Thus, these component will be integrated into TrOCR one by one. In this section, I will first mention the functionalities each component served in TrOCR, and then the three different integrated models will be introduced.

3.2.1 Recogniser - TrOCR

3.2.2 Corrector - CharBERT

3.2.3 Language Model - tbd

4 Experiment

4.1 Data

4.2 Ablation Study

4.3 Analysis

5 Results

5.1 Metric

Table 1 shows how to use the predefined tab command to have it listed.

language pair	ABC	YYY
EN→DE	20.56	32.53
DE→EN	43.35	52.53

Table 1: BLEU scores of different MT systems

And we can reference the large table in the appendix as Table 2

5.2 Evaluation

We saw in section **??**

We will see in subsection 5.2.1 some more evaluations.

5.2.1 More evaluation

5.3 Citations

Although BLEU scores should be taken with caution (see **?**) or if you prefer to cite like this: [**?**] ...

to cite: [**?**, 30-31]
to cite within parentheses/brackets: [**?**], [**?**, 30-32]

to cite within the text: ?, ?, 37
only the author(s): ?
only the year: ?

5.4 Graphics

To include a graphic that appears in the list of figures, use the predefined fig command:

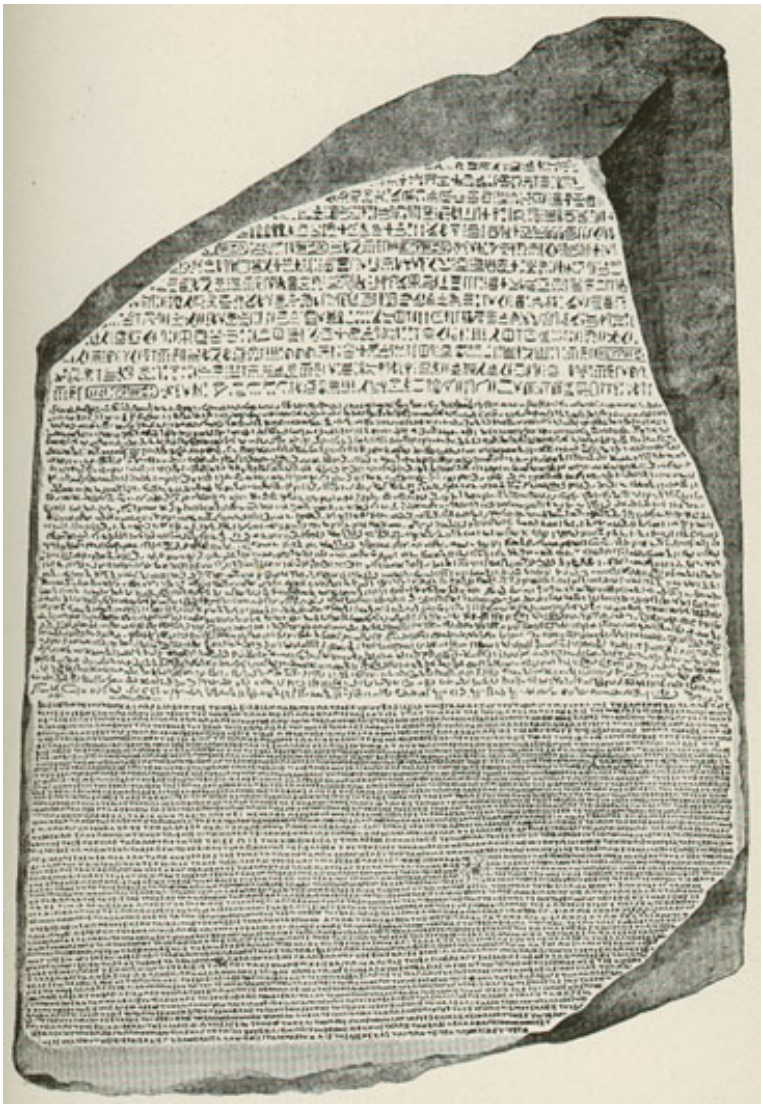

Figure 1: The Rosetta Stone

And then reference it as Figure 1 is easy.

5.5 Some Linguistics

(With the package 'covington')

Gloss:

(1) *The cat sits on the table.*
die Katze sitzt auf dem Tisch
'Die Katze sitzt auf dem Tisch.'

Gloss with morphology:

(2) *La gata duerm -e en la cama.*
Art.Fem.Sg Katze schlaf -3.Sg in Art.Fem.Sg Bett
'Die Katze schläft im Bett.'

6 Discussion

In this project we have done so much.[1]

We could show that ...

Future research is needed.
The show must go on.

[1] Thanks to many people that helped me.

7 Conclusion

In this project we have done so much.[1]

We could show that ...

7.1 Future Work

[1] Thanks to many people that helped me.

Glossary

Of course there are plenty of glossaries out there! One (not too serious) example is the online MT glossary of Kevin Knight [2] in which MT itself is defined as

> techniques for allowing construction workers and architects from all over the world to communicate better with each other so they can get back to work on that really tall tower.

accuracy A basic score for evaluating automatic **annotation tools** such as **parsers** or **part-of-speech taggers**. It is equal to the number of **tokens** correctly tagged, divided by the total number of tokens. [...]. (See **precision and recall**.)

clitic A morpheme that has the syntactic characteristics of a word, but is phonologically and lexically bound to another word, for example *n't* in the word *hasn't*. Possessive forms can also be clitics, e.g. The dog*'s* dinner. When **part-of-speech tagging** is carried out on a corpus, clitics are often separated from the word they are joined to.

[2] Machine Translation Glossary (Kevin Knight): http://www.isi.edu/natural-language/people/dvl.html

References

Curriculum vitae

Personal Information

Yung-Hsin Chen
yung-hsin.chen@uzh.ch

Education

2016-2020	Bachelor's degree in Physics at National Tsing-Hua University (NTHU)
since 2020	Master's degree in Informatics at University of Zurich (UZH)

Part-time Activities and Internships

2021	Delta Electronics Inc. Data Scientist Intern
2022-2023	Swiss Re Data Analyst

A Tables

Part of speech	POS type	POS	number of labels in my corpus
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	DET	**35**	280
14	Total	**35**	280

Table 2: Some very large table in the appendix

B List of something

This appendix contains a list of things I used for my work.

- apples
  - `export2someformat`
- bananas
- oranges
  - `bleu4orange`
  - `rouge2orange`