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

Supervisor: Dr. Phillip Ströbel, Dr. Simon Clematide

October 24, 2023

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

ACL	Association for Computational Linguistics
BNC	British National Corpus
CALL	Computer Assisted Language Learning
CFG	Context-Free Grammar
DG	Dependency Grammar
DTD	Document Type Definition
EACL	European Chapter of the Association for Computational Linguistics
MT	Machine Translation
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition
POS	Part-Of-Speech
TEI	Text Encoding Initiative
UTF-8	Unicode Transformation Format (8-bit)
XML	eXtensible Markup Language

1 Introduction

1.1 Motivation

Some words on your motivation would be nice.

1.2 Research Questions

The research questions that shall be answered in this thesis, are:

1. What do I do?
2. How do I do it?
3. And why?

1.3 Thesis Structure

In this first chapter ...
Chapter 2 introduces ...
Chapter 3 ...

2 Results

2.1 BLEU Scores

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

2.2 Evaluation

We saw in section 2.1

We will see in subsection 2.2.1 some more evaluations.

2.2.1 More evaluation

2.3 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: ?

2.4 Graphics

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

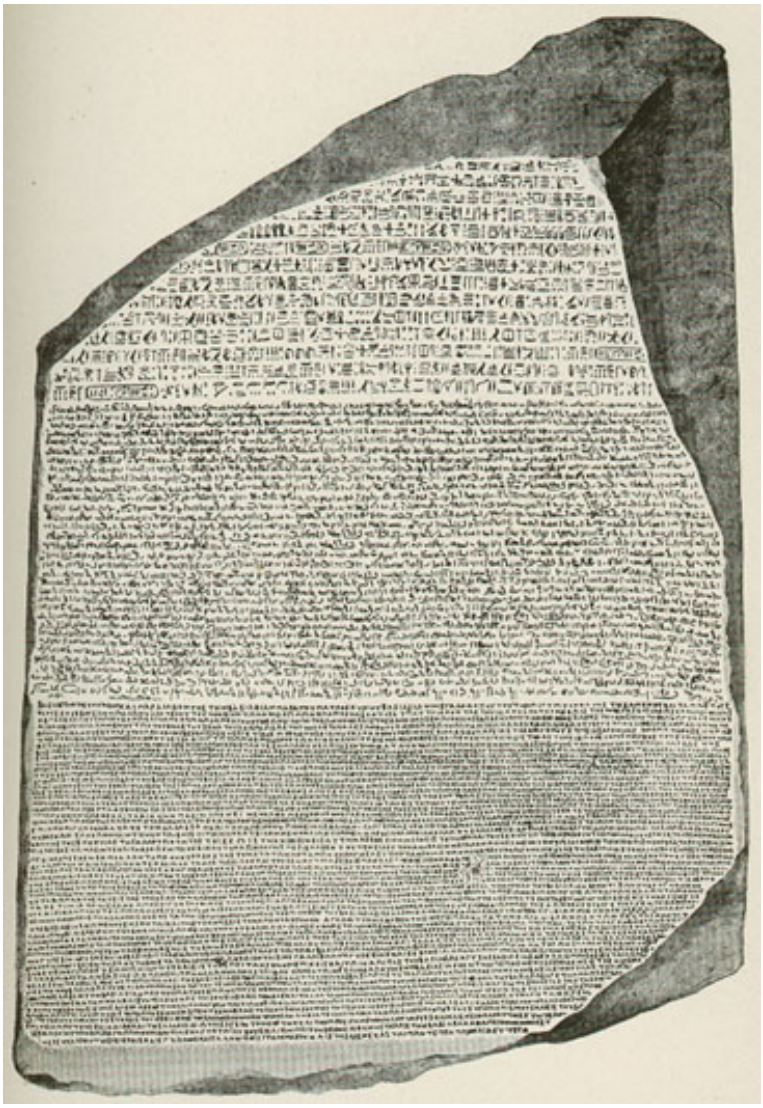

Figure 1: The Rosetta Stone

And then reference it as Figure 1 is easy.

2.5 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.'

3 Conclusion

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.

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

C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the European Association of Computational Linguistics (EACL)*, pages 249–256, 2006. URL `http://www.aclweb.org/anthology/E/E06/E06-1032.pdf`.

P. Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the MT Summit 2005*, pages 79–86, 2005. URL `http://www.iccs.inf.ed.ac.uk/~pkoehn/publications/europarl-mtsummit05.pdf`.

Lebenslauf

Persönliche Angaben

Ich Persönlich
Meinestrasse Nr
PLZ Wohnort
ichpersoenlich@uzh.ch

Schulbildung

2012-2014	Bachelor-Studium Computerlinguistik und Sprachtechnologie an der Universität Zürich
seit 2014	Master

Berufliche und nebenberufliche Tätigkeiten

2012–2013	Tutorate PCL I+II

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`