Glyph embedding for NMT

B-Town and K-Dog

https://github.com/lepidodendron/lepidodendron

Overview

1. Motivation
2. Implementation
3. Analysis
4. Future maneuvers

Wir begrüßen diese Entwicklung sehr.
We very much welcome this development.
We welcome this development.

Die Globalisierung hat unsere Industrien kaputt gemacht.
Globalism has destroyed our industries.
Globalisation has made our industries.

Das dürfte die meisten von uns nicht überraschen.
This is no surprise to most of us.
That should not be surprised by most of us.

Motivation

An alternative to learned embedding

- Problems with learned embedding
  - Vocabulary (over words, ngrams, or characters) is fixed and difficult to extend
  - Update frequencies for entries extremely unbalanced (Zipfian)
  - Can never model a language fully, even on the character level (2^{16} code points)
- Using glyphs as character embedding
  - Fixed data shape instead of fixed vocabulary size
  - Input space becomes continuous and open
  - Easier to visualize some choices made by model
  - Naturally suited for logographic languages
    - *火 fire 炎 fiery 灭 snuff 灰 ash 炭 coal*
- Glyph embedding for NMT
  - End-to-end image translation

Implementation

Dataset

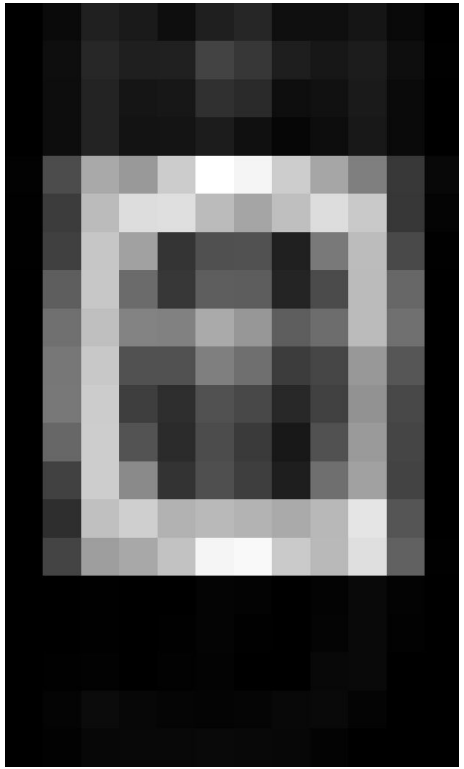
- European Parliament Proceedings Parallel Corpus 1996-2011
  - Paired corpus, not direct translation
- Experiments
  - Experiments conducted DE → EN
  - Took only sentences within [3,256] characters
  - 1 574 071 training instances
  - 4 096 validation instances

Europarl: A Parallel Corpus for Statistical Machine Translation, *Philipp Koehn, MT Summit 2005*

Converting strings to glyphs

- `noto sans mono` font, size 20, rendered with pillow
- height and width fixed by the largest frequent character
  - frequent characters are the ones that cover 99.95% of the texts
- pixels in grayscale [0,255], scaled to [0,1]
- ⍰ for *unk*, ▇ for *eos*, space for *bos*
  - *unk* is only relevant when we are not using glyphs

lang	charset	frequent chars	glyph dimension
DE	306	78	300 = 12x25
EN	293	72	240 = 12x20



Average English glyph

Rendered inputs

source

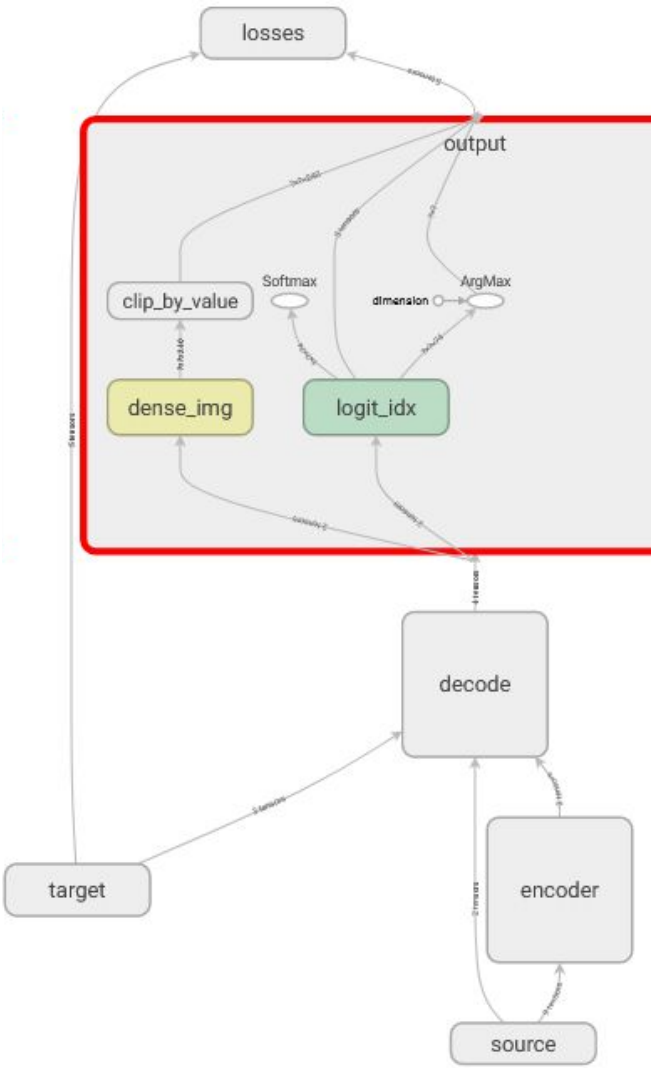
```
War dies nicht der Deal, den ich bereits 2005 prognostizierte?
Das wiederum ergibt im Spanischen responsabilidad democrática .
1. Iran: der Fall Roxana Saberi
Ihr Kampf um die Macht ist im Grunde ein Kampf um die Drogen.
Dies ist von entscheidender Wichtigkeit.
Nach der Tagesordnung folgt die Fragestunde (B5-0033/1999).
Immerhin steht Korruption in Kamerun auf der Tagesordnung.
Sie gründen im Kern in der Würde des Menschen.
```

target

```
Indeed, was this the deal that I predicted back in 2005?
In Spanish this gives, "responsabilidad democrática' .
1. Iran: the case of Roxana Saberi
Their fight for power is essentially a fight for drugs.
This is absolutely vital.
The next item is Question Time (B5-0033/1999).
Corruption is a fact of life in Cameroon.
They are, in essence, founded upon the idea of human dignity.
```

A standard nmt architecture

- RNN encoder
  - 3x stacked bidirectional GRU
- RNN decoder with attention
  - 3x causal GRU, followed by
  - multi-head scaled dot-product attention, with
    - residual connection
    - layer normalization
- hidden state units: 512
- dropout: 0.1
- adam with learning rate decay
- teacher-forcing training with batch size 128

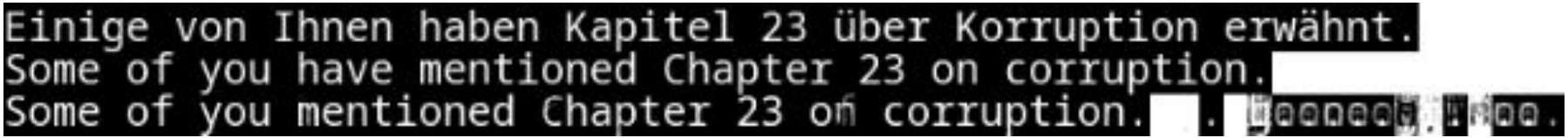

losses
output
clip_by_value
Softmax
dimension
ArgMax
dense_img
logit_idx
decode
target
encoder
source

Inputs and outputs

- **xyz** notation, **c** for character, **g** for glyph
  - **x**: encoder input
  - **y**: decoder input
  - **z**: decoder output
  - experiments: **ccc**, **cgc**, **cgg**, **ggg**
- decoder outputs
  - **z = c**: a dense layer after attention predicting chars, with softmax and XENT loss
  - **z = g**: a dense layer after attention predicting glyphs, with [0,1] clipping and MAE loss
  - whenever **z = g**, we also included a parallel char layer

Mismatch between glyph and char predictions

I agree.with thas.▉..▉.a.▉.bt...▉.....▉ust...▉.T.▉ .bel▉.sta.



Some of you mentioned Chapter 23 of corruption.▉ .▉ as.e. ▉S▉...

Mismatch between glyph and char predictions

Are there any comments?▉ a?be???an?ture?s??ook???▉u???r there?ar

Think of Vilvoordeaki▉▉use.is tuturaedn

Inference modes

mode	as autoregressive feedbacks	for bleu scores	z	y
1	g	argmax over c	g&c	g
2	g	g matched to char	g	g
3	g matched to char (and rendered)	g matched to char	g	c (g)
4	argmax over c (and rendered)	argmax over c	c	c (g)
5	probs over c (and rendered)	argmax over c	c	c (g)

- glyph-to-char matching (discretization) according to MAE
- inference terminates when the maximum step 256 is reached
- predicted strings are trimmed at the first *eos*

The 5 inference modes

Aber auch dieser Punkt ist noch offen.
But this too remains an open question.

But this issue is still open. uero. uo. oo. oo. oo ms also opon.

But this issue is still open. [redacted].ark. [redacted]ust [redacted]... [redacted] ak [redacted].a ms also apan.

But this issue is still open. uero. uo. oo. oo. oo ms also opon.

But this issue is still open. [redacted]uero. [redacted]ue. [redacted]oo. [redacted]uo. [redacted]oo as also opon.

But this is also open to this point. uoocia. uro. ust: this is a

But this is also open to this point. [redacted]uoocia. [redacted]uro. [redacted]ust: this is a

But this is also open. ubility. ust. . is also open. ust. . is

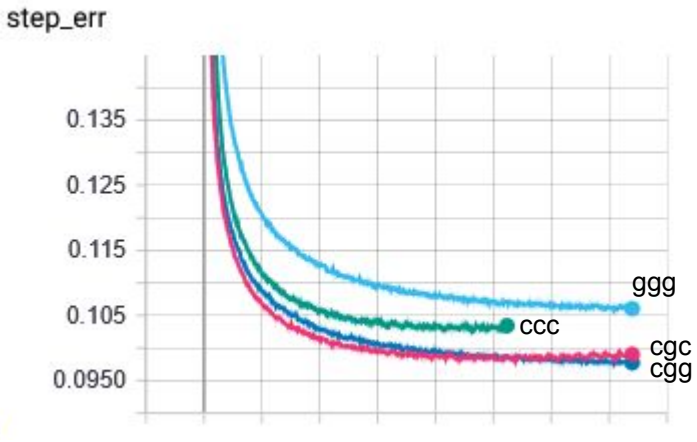
But this is also open. [redacted]ubility. [redacted]ust. [redacted]. [redacted] is also open. [redacted]ust. [redacted]. [redacted] is

Best ieteraty, toolegene is anounit. U. U. , C. torn again. .

Best ieteraty, toolegene is anounit. [redacted]..[redacted]..[redacted]. C. torn again. [redacted].[redacted]

Results

- `sacrebleu -tok intl`
- **ccc** overfit early



model	epochs	mode 1	mode 2	mode 3	mode 4	mode 5
ccc	30	n/a	n/a	n/a	**30.9**	20.3
cgc	60	n/a	n/a	n/a	**30.9**	22.4
cgg	60	**22.9**	**23.5**	**26.5**	**30.9**	**24.1**
ggg	60	21.9	22.2	25.3	30.2	23.0

Analysis

Decision points

Freizügigkeit für Personen
Free movement of persons
Freedom of movement for people[illegible] freedom[illegible]
Free movement of persons[illegible]

cgg: Free om of movement for people[illegible] free om on migpeers
ggg: Free movement of persons[illegible]y brto ovc of tucomaa o.[illegible].ortei [illegible].[illegible]([illegible]

Wir begrüßen diese Entwicklung sehr.
We very much welcome this development.
We welcome this development.[illegible]
We very much welcome this development.[illegible]

cgg: We velcome this development.[illegible] SS werkine.[illegible] O wncurs toisgest[illegible].s
ggg: We very much welcome this development.[illegible] of these.[illegible] eth [illegible]uels.[illegible].[illegible]

Decision points

Das ist jedoch nicht geschehen.
This is not, however, what has happened.
That has not happened, however.
That has not been dotooned.

cgg: That has not happened. however. Nesn. ustir. andenl. Tuthuld.
ggg: That has not heen drtammed. .af the sucomda. Aask Aitomis. hot i

Das dürfte die meisten von uns nicht überraschen.
This is no surprise to most of us.
That should not be surprised by most of us.
That should not be surprised by most of us.

cgg: That itdris should not ce surprised.ay most.of us. .asbod. au t
ggg: That stvses should not be surprised by most of us. . .er an tun

Imagined words

Aber diese Denkungsart ist völlig falsch!
How misguided we are!
But this manoul is completely wrong. [illegible]
But this is completely wrong. [illegible] the [illegible]

cgg: But this manoul is completely wrong.[illegible]n aurr[illegible]
ggg: But this ms completely wrong[illegible]ueti[illegible]uasti[illegible]usts p the fttestin

Aber Sie werden sie bekommen.
Rest assured that you will receive a reply in due course.
But you will [illegible] them. [illegible]
But you will [illegible] them. [illegible] of them. [illegible] of the [illegible]

cgg: But you will brcopp them.[illegible]...n....[illegible]...[illegible].esn.[illegible]...[illegible].asn Cner.yu.[illegible].
ggg: But you will gnli e them.[illegible]uetle.[illegible].tunlens.[illegible] if them.[illegible].tf the Uuc

Synonym struggles

Das ist unangenehm.
That is embarrassing.
That is unclared. [illegible] is uncecembul. [illegible]
This is anpleased. [illegible] of tonts. [illegible] the Etturity. [illegible] t

cgg: That is unclased.[illegible]uTE[illegible]H is uncecembul..[illegible].a[illegible]..[illegible].[illegible] T[illegible]
ggg: This is anpleasid.[illegible] if tonts.[illegible].titkes.[illegible].if the Etturity.[illegible].[illegible].if t

Das ist ein wichtiger Hinweis.
It is important that that should be said.
This is an important [illegible] [illegible]
This is an important comment. [illegible] to be [illegible] ual. [illegible] o

cgg: This is an important polter.[illegible]unt[illegible].a.[illegible].at[illegible]urt[illegible]uamr [illegible]usti [illegible]B2 rnde
ggg: This is an important plmment.[illegible].eswiudh.to te ane[illegible]uali[illegible]ient.[illegible]utnm

Synonym struggles

Dies ist von entscheidender Wichtigkeit.
This is absolutely vital.
This is critical.
This is crucial.

cgg: This is vriticall stunt.. .. . . uet .e. ua is auveful. .s.
ggg: This is crucial. ieds. wutkrios. if tuttoe .ann.and ihe Eutult

Wir haben einen völlig unwahren Text angenommen.
The text which we have approved is totally wrong.
We have adopted a completely untrabteated text.
We have adopted a completely untransable text.

cgg: We have adopted a tompletely untrubteated text. aspea... .a....
ggg: We have adopted a completely untransable text. uats uslt uptu t

Is this the end?

Die Kommission reagierte sofort.
The Commission reacted very promptly.
The Commission responded immediately [illegible]
The Commission immediately responded to the [illegible] of tha

cgg: The Commission reaponded immediately.wint ne...[illegible].[illegible]...[illegible].ezn.[illegible].[illegible]
ggg: The Commission rmmediately responded to the nttenvattat.[illegible].if tha

Ich hoffe sehr, dass wir ein ...
I hope very much that we will maintain a ...
I very much hope that we [illegible]
I very much hope that we [illegible] the [illegible]

cgg: I very much hope that we wnderl ..t .n u ....[illegible].[illegible]....chrk. [illegible] e [illegible]
ggg: I very much hope that we .neur..[illegible]f.u....li....I .lsosee..he....s

Is this the end?

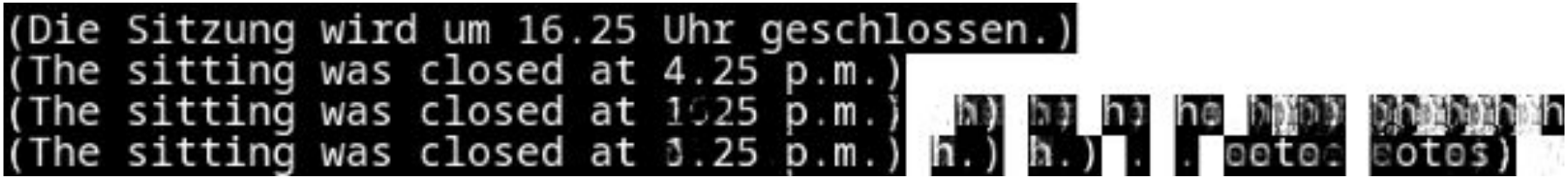
Das ist die eine Seite.
That is what is happening on the one hand.
That is one side of the [illegible]

cgg: That is one side.of the ppc....[illegible].s.[illegible]..[illegible]urt.[illegible]urt[illegible].[illegible]rianesrth [illegible]ust

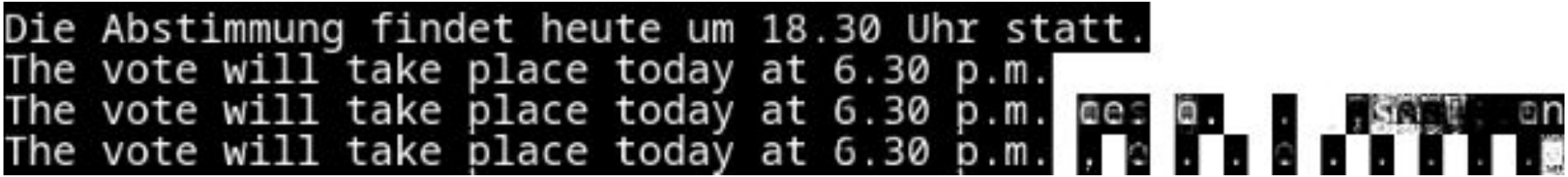
Es ist auch leicht zu erklären, warum.
It is also easy to explain why.
It is also easy to explain why [illegible]

cgg: It is also easy to explain why.toou.iusk.[illegible].[illegible] in in turt [illegible]u an

Hour format



cgg: (The sitting was closed at 4.25 p.m.)▇▇.)▇..▇h.▇h.▇D▇C.▇ChCCChCh
ggg: (The sitting was closed at 4.25 p.m.)▇h.)▇h.)▇.▇.▇.ete-▇.otes)▇▇



cgg: The vote will take place today at 6.30 p.m.▇.a.▇..▇▇.▇▇ ▇ CRwtn
ggg: The vote will take place today at 6.30 p.m.▇.▇.▇.▇.▇.▇.▇.▇.▇.(

Unknown chars get correct glyphs

Bericht Theato (A5-0090/1999)
Theato report (A5-0090/1999)
Theato report (A5-0090/1999)

cgg: Theato report (A5-0090⁄1999)

Anfragen an Herrn Byrne
Questions to Commissioner Byrne
Questions to Mr Byrne

cgg: ⁄uestions to Cr Byrne

Simbabwe war einst ein blühendes Land und könnte es wieder sein.
Zimbabwe was once a flourishing country and it could be again.
Zimbabwe was once a pronouncing country and could be reserved a

cgg: ⁄imbabwe was once a pron ri ging country and could be reptrved.a

Recover from noisy predictions

Gemeinsamer Entschließungsantrag
Joint motion for a resolution
Gemeinsamer Entschließungsantrag motion for a resolution

cgg: Joneriona motion for a resolution J

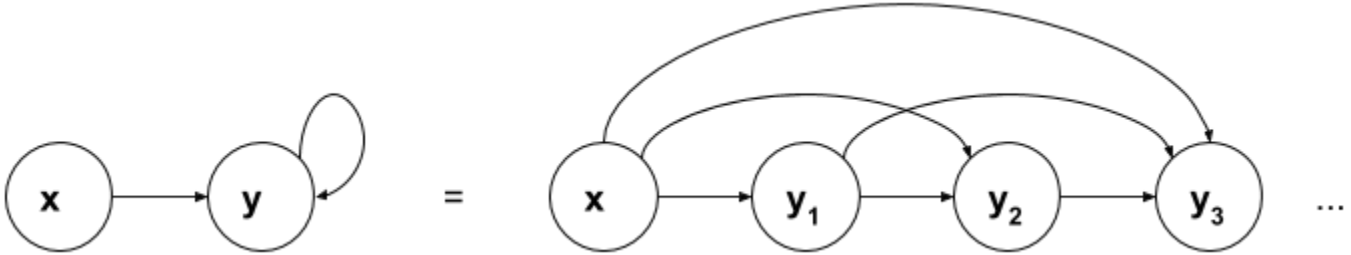
Dies wäre meines Erachtens falsch.
That would seem wrong to me.
I thinkthis would be wrong.

cgg: I tolr this would be wrong. asbed.. e. ..Nat s (ooniuldr

Er ist in meinen Augen auch der wichtigste.
In my eyes this is the most important.
I meant that it is the most important in my view.

cgg: I tett to k that it is the most important.oucn.in my view.uster

Excursion: autoregressive seq2seq models



- $p(y \mid x) = p(y_1 \mid x)\, p(y_2 \mid y_1, x)\, p(y_3 \mid y_2, y_1, x) \ldots$
  - Each y_n is a discrete random variable over the target vocabulary T
  - The whole search space is T^n
- But glyphs are not predicted as a random variable
  - A glyph-predicting decoder is not truly autoregressive
  - Unless we perform discretization (mode 3)

A second look at the results

- Stochastic modeling works better (modes 3&4), however
  - fuzzy glyphs are less problematic than fuzzy chars (mode 5)
- The decoder benefits from using glyphs
  - a non-autoregressive model with convolution may benefit more

model	mode 1	mode 2	mode 3	mode 4	mode 5
ccc	n/a	n/a	n/a	30.9	20.3
cgc	n/a	n/a	n/a	30.9	22.4
cgg	22.9	23.5	26.5	30.9	24.1
ggg	21.9	22.2	25.3	30.2	23.0

Future maneuvers

采薇采薇
靡室靡家

采薇采薇
憂心烈烈

采薇采薇
王事靡盬

彼爾維何
戎車既駕

駕彼四牡
四牡翼翼

昔我往矣
行道遲遲

薇亦作止
獫狁之故

薇亦柔止
載飢載渴

薇亦剛止
不遑啟處

四維常之華
四牡業業

四牡騤騤
象弭魚服

楊柳依依
載渴載飢

日歸日歸
不遑啟居

日歸日歸
我戍未定

日歸日歸
憂心孔疚

彼路斯何
豈敢定居

君子所依
豈不日戒

今我來思
我心傷悲

歲亦莫止
獫狁之故

心亦憂止
靡使歸聘

歲亦陽止
我行不來

君子之車
一月三捷

小人所腓
獫狁孔棘

雨雪霏霏
莫知我哀

Logographic language translation

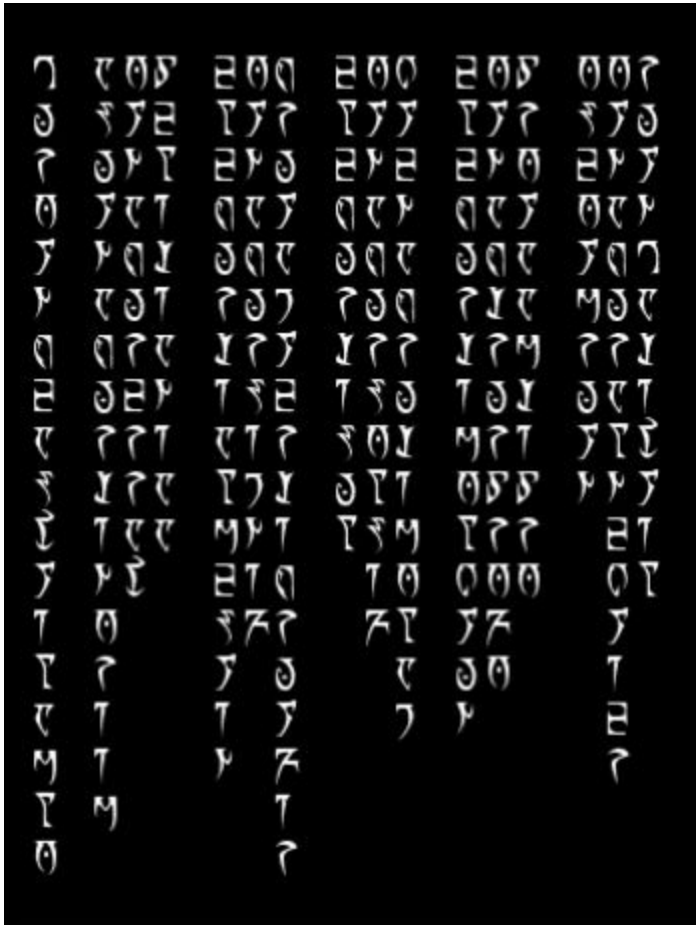
- Chinese to English translation with UNCorpus
- Experiments
  - We take only sentences within [3,128] characters
  - 7 392 227 training instances
  - 4 096 validation instances
  - To compare **cgg** and **ggg**

lang	charset	frequent chars	glyph dimension
ZH	6357	2681	500 = 20x25
EN	714	80	240 = 12x20

Directions for future works

- CNN
  - convolution over rendered image
  - transposed convolution for non-autoregressive prediction
- Without character boundaries
  - simply an image instead of a sequence of glyphs
  - postnet for word (piece) prediction with CTC loss
- Multiple fonts and typefaces
  - combinatorially more training data

new
meaning



old
form