The Same is Not The Same

Postcorrection of Alphabet Confusion Errors in Mixed-Alphabet OCR Recognition

by Jonas Hempel

Bulgerian-German to English

Иван ора нивата

In English: Ivan plowed the field.

'opa' is German word for 'grandfather'

Alphabet Similarities (1)

Latin	*A*	*B*	*C*	*E*	*H*	*K*	*M*	*O*	*P*	*T*	*X*	*Y*	*a*	*c*	*e*	*g*	*k*	*m*	*n*	*o*	*p*	*u*	*x*	*y*
Cyrillic	*A*	*B*	*C*	*E*	*H*	*K*	*M*	*O*	*P*	*T*	*X*	*У*	*a*	*c*	*e*	-	-	*m*	*n*	*o*	*p*	*u*	*x*	*y*
Latin	A	B	C	E	H	K	M	O	P	T	X	Y	a	c	e	g	k	m	n	o	p	u	x	y
Cyrillic	A	B	C	E	H	K	M	O	P	T	X	У	a	c	e	g	k	m	n	o	p	u	x	y

- Latin-Cyrillic transition table
- Upper font is Times New Roman
- Lower font is Universum

Table taken from the paper.

Alphabet Similarities (2)

Latin	*A*	*B*	*E*	*Z*	*H*	*I*	*K*	*M*	*N*	*O*	*P*	*T*	*Y*	*a*	*y*	*n*	*i*	*o*	*p*	*u*	*w*
Greek	*A*	*B*	*E*	*Z*	*H*	*I*	*K*	*M*	*N*	*O*	*P*	*T*	*Y*	*α*	*γ*	*η*	*ι*	*o*	*ρ*	*υ*	*ω*
Latin	*A*	*B*	*E*	*Z*	*H*	*I*	*K*	*M*	*N*	*O*	*P*	*T*	*Y*	*a*	*y*	*n*	*i*	*o*	*p*	*u*	*w*
Greek	*A*	*B*	*E*	*Z*	*H*	*I*	*K*	*M*	*N*	*O*	*P*	*T*	*Y*	*α*	*γ*	*η*	*ι*	*o*	*ρ*	*υ*	*ω*

- Latin-Greek transition table
- Upper font is Times New Roman
- Lower font is Verdana Cursive

Table taken from the paper.

Training and Test corpora

- Sophia-Munich corpus
- Bulgarian EC corpus
- Greek-Latin corpus

Algorithm

- Levenshtein distance $d_0(w_i, v)$
- Normalized similarity value $s(v, w_i)$
- collocation frequency value $f(v, w_{i-1}, w_{i+1})$

$\rightarrow$ $score(v) = \alpha * s(v, wi) + (1-\alpha) * f(v)$

- α balance parameter
- τ threshold parameter

Evaluation Results (1)

Corpus/OCR	tokens	error rate OCR → pc	ac-error rate OCR → pc
SM-OCR1-Tr	8110	11.22→ 6.57%	5.42→ 1.25%
SM-OCR1-Te	7923	10.59→ 6.25%	5.44→ 1.26%
SM-OCR2-Tr	5099	37.24→ 15.87%	9.96→ 1.75%
SM-OCR2-Te	5115	43.63→ 16.81%	10.28→ 2.01%
EC-OCR1-Tr	6571	15.05→ 5.68%	10.94→ 1.81%
EC-OCR1-Te	6230	16.44→ 7.03%	13.00→ 3.74%
EC-OCR2-Tr	6571	48.52→ 9.0%	27.71→ 4.14%
EC-OCR2-Te	6230	48.81→ 11.35%	27.50→ 3.23%

- Bulgarian Sophia-Munich and Bulgarian EC corpus
- Error rate for plain OCR recognition and postcorrection
- Training (Tr) and Test (Te) data
- ac-error: alphabet confusion error

Table taken from the paper.

Evaluation Results (2)

Corpus/OCR	err.r.	ac-errors	ac-err. r.
GR-OCR1-Ti	2.37%	62	0.74%
GR-OCR1-Vd	2.12%	49	0.58%
GR-OCR2-Ti	12.00%	550	6.54%
GR-OCR2-Vd	10.87%	457	5.43%

- Greek newspaper corpus
- Cursive Times (Ti) and cursive Verdana (Vd) font
- ac-error: alphabet confusion error

Table taken from the paper.