Poor Man's OCR Post-Correction: Unsupervised Recognition of Variant Spelling Applied to a Multilingual Document Collection

Harald Hammarström
Uppsala University
Shafqat Virk and Markus Forsberg
Språkbanken, Gothenburg University

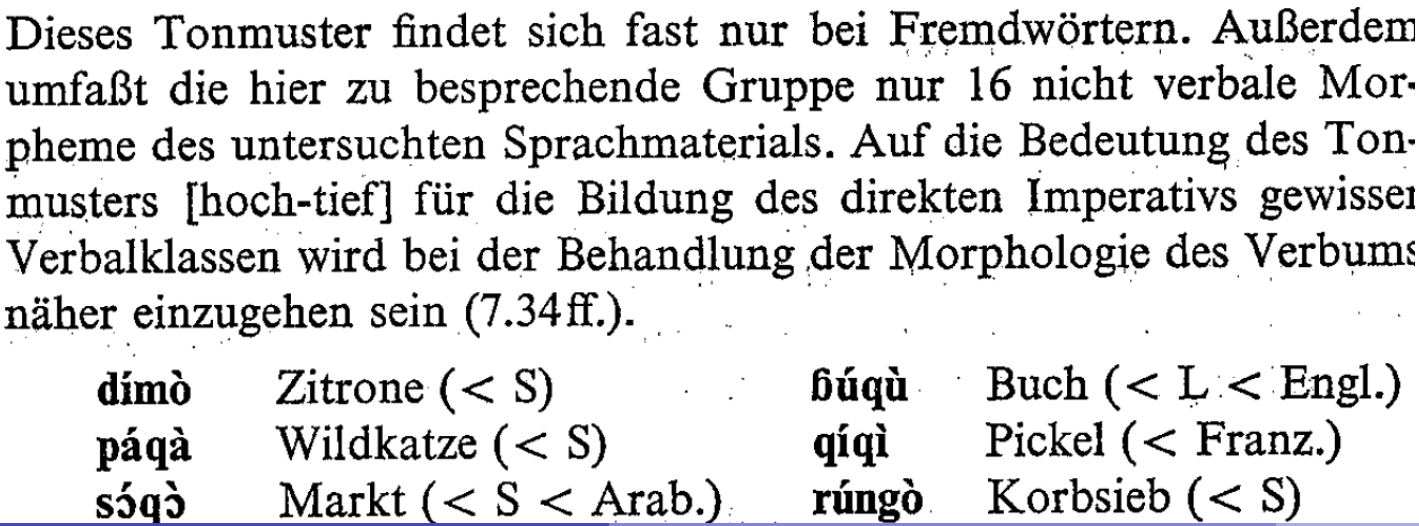
1-2 June 2017 Göttingen

Poor Man's OCR Post-Correction: Motivation

- We have an OCRed document collection of descriptive grammars
  - Spans a dozens of different (meta-)languages
  - OCR quality quite varied
  - Important genre-specific terms
- Existing OCR post-correction techniques require
  - Resources (language-specific)
  - Tuning and adaptation
- A light-weight genre- and language-independent approach needed (even if not state-of-the-art accuracy for English)

OCR Quality Example (Though Quality Varies)

```
Dieses Tonmuster findet sich fast nur bei Fremdwörtern. Au$$erdem
umfa$$t die hier zu besprechende Gruppe m1r 16 nicht verbale
Morpheme des untersuchten Sprachmaterials. Auf die Bedeutun& des
Tonmus.ters [hoch-tief] für die Bildung des direkten Imperativs gew
Verbalklassen wird bei der Behandlung .der Morphologie des Verbums
nähereinzugehen sein (7.34ff.). Âů
Âů
Âů
dimo
paqa
s~q;,
```



Dieses Tonmuster findet sich fast nur bei Fremdwörtern. Außerdem umfaßt die hier zu besprechende Gruppe nur 16 nicht verbale Morpheme des untersuchten Sprachmaterials. Auf die Bedeutung des Tonmusters [hoch-tief] für die Bildung des direkten Imperativs gewisser Verbalklassen wird bei der Behandlung der Morphologie des Verbums näher einzugehen sein (7.34ff.).

dímò	Zitrone (< S)	**ɓúqù**	Buch (< L < Engl.)
páqà	Wildkatze (< S)	**qíqì**	Pickel (< Franz.)
sóqò	Markt (< S < Arab.)	**rúngò**	Korbsieb (< S)

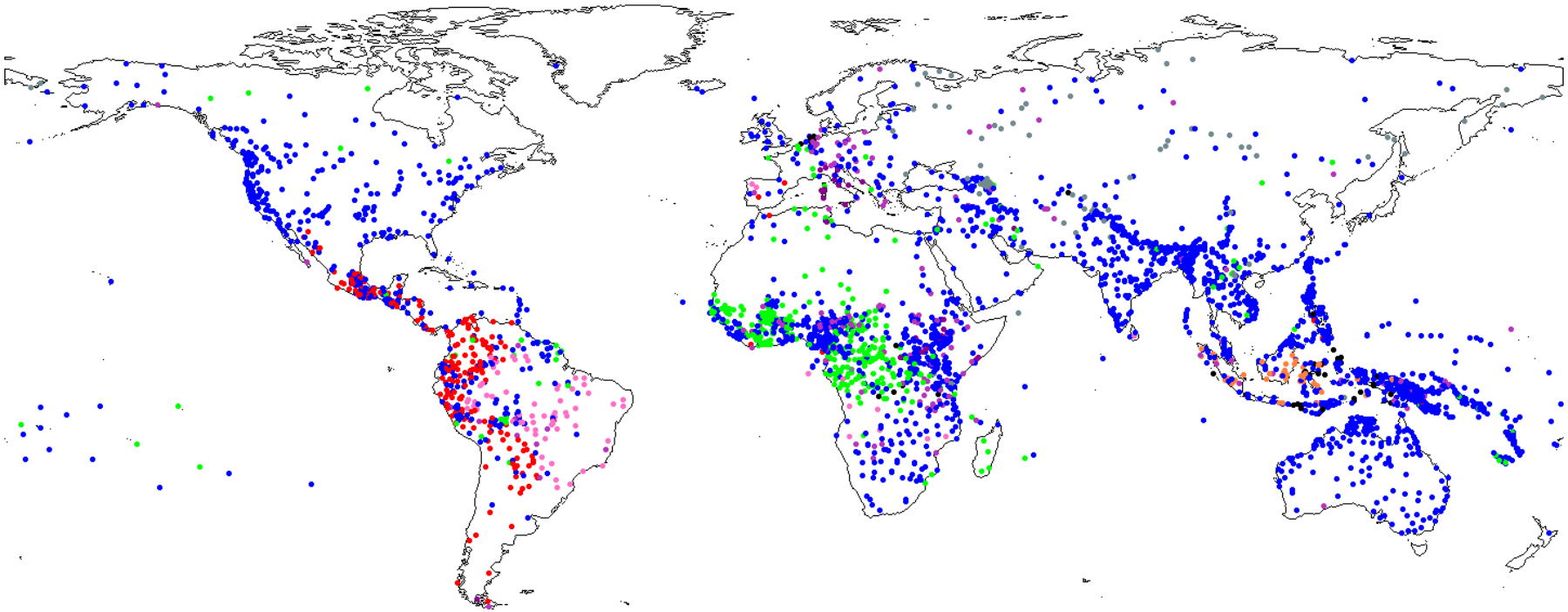
OCRed Grammar Collection

Spans 4 005 (target-)languages written in 96 (meta-)languages

Meta-language		# lgs	# Doc:s	# Types	# Tokens
English	eng	3098	23 708	23 114 708	380 467 360
French	fra	680	3 452	3 585 529	86 699 512
German	deu	468	2 753	2 830 285	38 643 792
Spanish	spa	332	2 484	2 490 063	84 925 065
Portuguese	por	115	1 076	716 078	14 420 655
Russian	rus	220	677	293 909	50 387 961
(Mandarin	cmn	94	445	?	?)
Dutch	nld	74	528	397 564	5 849 144
Italian	ita	64	329	555 043	6 058 028
Indonesian	ind	70	206	166 524	2 163 114
...	...	...	...	...	...

English accounts for a larger share than all the other ones together!

Geographical Distribution of Meta-Languages



eng	blue	deu	purple	ind	orange
fra	green	rus	slate gray	nld	black
spa	red	por	neonpink	ita	magenta

OCR quality is a significant issue!

- Probably deep-parse of the texts are is not feasible for a fair share of the documents
- OCR errors affect genre-specific highly important terms, e.g., terms distributionally similar to `SOV`

term	Distributional similarity to 'SOV'
SVO	0.96
VSO	0.90
SOY	0.89
VOS	0.83
AOV	0.80
...	...

- Correct OCR of terms of vernular (described) languages not feasible and not targeted

OCR Post-Correction State-of-the-Art

- Similar to spelling correction, going back to Damerau (1964)

  Correct an out-of-dictionary word to a dictionary word that is similar in form (Eger et al., 2016)

- Features based on form, frequency and dictionary properties used to rank candidates (Evershed and Fitch, 2014)
- Most systems use labeled training data and handle it as a regular supervised Machine Learning problem (Mei et al., 2016, Reffle and Ringlstetter, 2013, Silfverberg et al., 2016)
  - Though Afli et al. (2016), Afli and Way (2016) treat it as an SMT problem (thereby making some use of context)
- A few systems use context explicitly (Evershed and Fitch, 2014, Tong and Evans, 1996)

All systems rely on a **dictionary** *of correct forms and/or* **threshold** *tuning*

Poor Man's OCR Correction: Principles

- No dictionary, no labeled training data, no thresholds
- OCR corrects types, so requires that the orthography of the input language has word boundaries

1. $Sim(x, y)$: From raw text data get the *distributional similarity* between terms x, y (using a small size window in Word2Vec, Mikolov et al. 2013)
2. $N(x) = \{y | ED(x, y) \leq 1\}$: The form-neighbourhood, giving the set of forms close to x
3. $V(x, y)$: y is an OCR variant of x iff $S(x, y)$ exceeds that expected by chance from $|N(x)|$ random trials
4. $V(x, y)$ would be sufficient except any language might also have true minimal pairs, so also check if the relative frequencies of x vs y resemble that of OCR errors rather than minimal pairs

Example: Distributional Similarity for 'language'

The term 'language' has a distributional similarity to every other of 204 002 word types

Rank	y	$S(language, y)$
1	languages	0.7619
2	linguistic	0.7555
3	dialect	0.7381
4	community	0.7074
5	history	0.7036
6	culture	0.6995
7	society	0.6704
8	population	0.6636
9	lexicon	0.6542
10	literature	0.6482
...	...	...
100	quiche	0.5584
...	...	...
100000	diversa	0.0269
...	...	...

Example: Form Neighbourhood of 'language'

$$N(\textit{language}) = \{\textit{aanguage}, \textit{banguage}, \ldots, \textit{zanguage}\ \{\textit{language}, \textit{lbnguage}, \ldots, \textit{lznguage}\ \ldots, \ldots, \ldots\ \textit{languaga}, \textit{languagb}, \ldots, \textit{languagz}\}$$

- Contains $|\textit{language}| \cdot |\Sigma| = 8 \cdot 26 = 208$ forms (if Σ is the English lowercase alphabet)
- Which of these 208 forms have a higher than expected distributional similarity $S(\textit{language}, y)$ to 'language'?
- If you draw k out of n values the expectation is that the maximum of the k values is at the $k+1$th quantile
- The total vocabulary size is 204 002 and on $|N(\textit{language})| = 208$ trials the expected quantile to beat is the $\frac{1}{209} \cdot 204002 \approx 976$th quantile
- The 976th highest value of $Sim_{\textit{language}}(y)$ is 0.2839

Example: Similarity and Form Neighbours of 'language'

- 7 of the members of $N(language)$ have a distributional similarity to 'language' higher than 0.2839

y	$Sim(x,y)$	$f(y)$	$\frac{f(y)}{f(x)+f(y)}$	$\frac{f(y)}{f(x)+f(y)}/S(x,y)$
ianguage	0.52356	387	0.00066	0.00125
languagc	0.50100	225	0.00038	0.00076
languaqe	0.44455	93	0.00016	0.00035
languuge	0.29799	68	0.00012	0.00038
languago	0.37767	135	0.00023	0.00060
lauguage	0.34320	77	0.00013	0.00038
lunguage	0.46430	63	0.00011	0.00023

- Those terms are deemed OCR variants (of 'language') whose frequency $f(language) = 581815$ is much higher

But what about (true) minimal pairs?

- A natural language may have forms that are minimal pairs that happen to be similar distributionally, and some of those with high token frequency, e.g., English *in* and *on*
  - The poor man's approach (so far) will think the less frequent one is an OCR error for the other
- In most OCR post-correction systems, such corrections are avoided by recourse to the dictionary (which will whitelist both forms)
- Instead of a dictionary, the poor man can use the following heuristic
  - If y really is an OCR error for x then its frequency should be derived from x's frequency at some error rate r
  - We do not know r but can estimate an upper bound on the rate by looking at **all** pairs in $V(x, y)$ (the real OCR errors plus the minimal pairs)
  - If the frequency of y relative to x scaled by $S(x, y)$ (the extent to which they occur in the same circumstances) is so high that it surpasses even this rate, it is not believable that it is derived solely from faulty occurrences of x
- In the present test set $r_V = \frac{458626300}{2714497206} \approx 0.16895$.

Example: OCR variants vs minimal pairs

y	$Sim(x,y)$	$f(y)$	$\frac{f(y)}{f(x)+f(y)}$	$\frac{f(y)}{f(x)+f(y)}/S(x,y)$
then	0.51802	411360	0.25513	**0.49250**
them	0.70800	378516	0.23964	**0.33847**
thcy	0.32272	1256	0.00104	0.00323
thoy	0.42350	1760	0.00146	0.00345
thej	0.29713	292	0.00024	0.00081
theg	0.33179	143	0.00012	0.00035
ihey	0.42526	882	0.00073	0.00172
thev	0.43283	822	0.00068	0.00158
tney	0.29813	174	0.00014	0.00048
tbey	0.39003	1210	0.00101	0.00258
fhey	0.35574	129	0.00011	0.00030

It is not believable that 'then'/'them' are OCR errors for 'they' because they correspond to error rates of 0.49/0.33, far greater than even the upper bound $r \approx 0.16895$

Evaluation of Poor Man's OCR Post-Correction

Evaluation of poor man's OCR post-correction on the Sydney Morning Herald 1842-1954 dataset 2 (Evershed and Fitch, 2014).

	Dataset 2		After OCR correction		
	Types	Tokens		Types	Tokens
	11 650	38 226		11 650	38 226
Correct forms	7 655	32 714	untouched	7 383	32 225
			hypercorr.	272	489
Erroneous forms	3 995	5 512	untouched	3 152	4 276
			corrected	540	874
			adjusted	303	362

Word Error Rate improves from 85.5% to 86.5% (though this is significantly lower than the 93.7% achieved by Evershed and Fitch 2014)

Type reduction after OCR post-correction

Meta-language		# Types Before	After	Reduction
English	eng	23 114 708	21 681 596	6.2%
French	fra	3 585 529	3 399 081	5.2%
German	deu	2 830 285	2 742 546	3.1%
Spanish	spa	2 490 063	2 373 030	4.7%
Portuguese	por	716 078	707 583	1.2%
Russian	rus	293 909	293 370	0.2%
Dutch	nld	397 564	394 171	0.9%
Italian	ita	555 043	550 359	0.8%
Indonesian	ind	166 524	164 524	1.2%

Conclusion and Outlook

- OCR correction off-the-shelf: no dictionary, no labeled training data, no thresholds, no tuning, ...
- Accuracy nevertheless lower than methods which make use of resources
- The poor man's OCR correction method may be iterated
- Open-source Python implementation (less than 50 lines of code)

  https://github.com/shafqatvirk/contentProfiling/ocrc.py

Afli, H., Qiu, Z., Way, A., and Sheridan, P. (2016). Using SMT for OCR error correction of historical texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, pages 962–966, Portorož, Slovenia. ELRA.

Afli, H. and Way, A. (2016). Integrating optical character recognition and machine translation of historical documents. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 109–116, Osaka, Japan. The COLING 2016 Organizing Committee.

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Eger, S., vor der Brück, T., and Mehler, A. (2016). Statistical learning for OCR text correction. *The Prague Bulletin of Mathematical Linguistics*, 105:77–99.

Evershed, J. and Fitch, K. (2014). Correcting noisy ocr: Context beats confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 45–51, New York, NY, USA. ACM.

Mei, J., Islam, A., Wu, Y., Moh'd, A., and Milios, E. E. (2016). Statistical learning for OCR text correction. *CoRR*, abs/1611.06950:1–10.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 3111–3119. Neural Information Processing Systems, Lake Tahoe, Nevada.

Reffle, U. and Ringlstetter, C. (2013). Unsupervised profiling of ocred historical documents. *Pattern Recognition*, 46(5):1346–1357.

Silfverberg, M., Kauppinen, P., and Lindén, K. (2016). Data-driven spelling correction using weighted finite-state methods. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 51–59, Berlin, Germany. Association for Computational Linguistics.

Tong, X. and Evans, D. A. (1996). A statistical approach to automatic ocr error correction in context. In *Fourth Workshop on Very Large Corpora*, pages 88–100, Copenhagen. Association for Computational Linguistics.