A Taxonomy of Weeds: A Field Guide for Corpus Curators to Winnowing the Parallel Text Harvest

Katherine M. Young Jeremy Gwinnup Lane O. Schwartz

N-Space Analysis, LLC
katherine.young.ctr.1@us.af.mil

Air Force Research Laboratory
jeremy.gwinnup.1@us.af.mil

University of Illinois at Urbana-Champaign
lanes@illinois.edu

31 October 2016

Modern MT requires corpora

Weeds among the wheat

"Weed is possibly a better metaphor than dirt for many of these issues with SMT training data: a wild plant growing where it is not wanted and in competition with cultivated plants"

– Simard (2014)

Weeds among the wheat

ἐν δὲ τῷ καθεύδειν τοὺς ἀνθρώπους ἦλθεν αὐτοῦ ὁ ἐχθρὸς καὶ ἐπέσπειρεν **ζιζάνια** ἀνὰ μέσον τοῦ σίτου καὶ ἀπῆλθεν.

Parable of the weeds
—Greek New Testament

Some types of weeds

zizania ex machina

- Wrong language text
- Historical encoding errors
- Bidirectional reversal
- Sentence alignment errors
- Sentence-internal fragments
- Harvested MT

zizania ex homine

- Mixed alphabet spellings
- Mixed morphology
- Transliterations of NE and Borrowings
- Underachieving translation
- Overachieving translation
- Translation directionality

Weeds of mechanical origin

Wrong language text

IWSLT 2014

Wrong language text

IWSLT 2014

Wrong language text

Common Crawl

Wrong language text

Common Crawl

Wrong language text

- Ukrainian I (i I)
- Yi(ï Ï)
- Ghe With Upturn (ґ Ґ)
- Ie (є Є).

Wrong language text

Wrong language text

+0.34 BLEU on WMT’15.

Historical encoding errors

Ñïðàâêà ïî ãîðîäàì Ðîññèè è ìèðà

Historical encoding errors

Ñïðàâêà ïî ãîðîäàì Ðîññèè è ìèðà

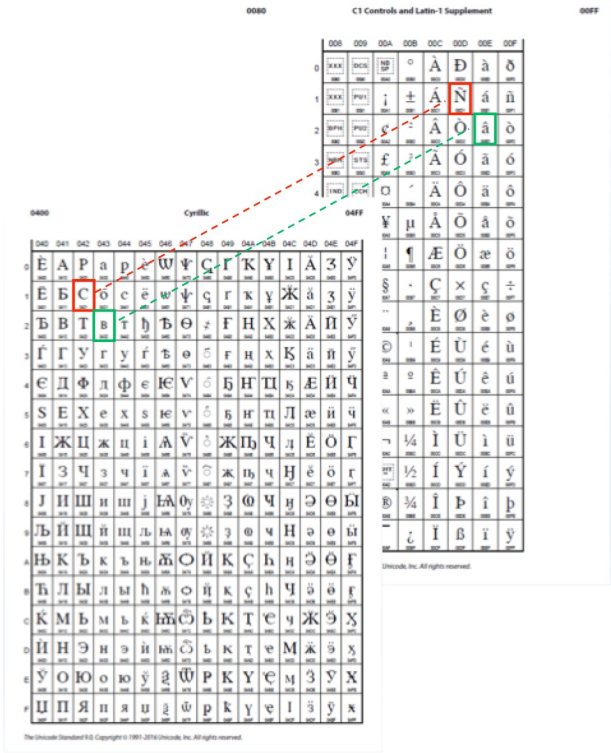
Справка по городам России и мира

Historical encoding errors

In ru-en common crawl, some Russian text is actually encoded as Windows-1251 but is incorrectly interpreted as UTF-8

Ñïðàâêà ïî ãîðîäàì Ðîññèè è ìèðà.

Справка по городам России и мира.



Correct by shifting 350_{hex} code points into Unicode Cyrillic range

Historical encoding errors

йquipe chʙteau

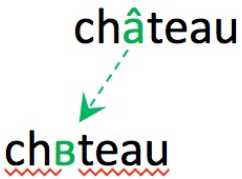
Historical encoding errors

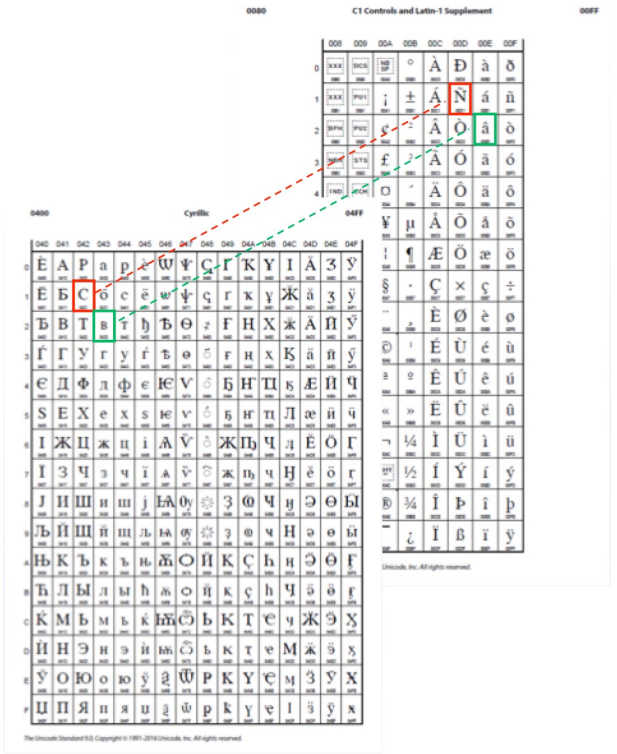
йquipe chʙteau

équipe château

Historical encoding errors

In fr-en common crawl, some French text has the reverse problem

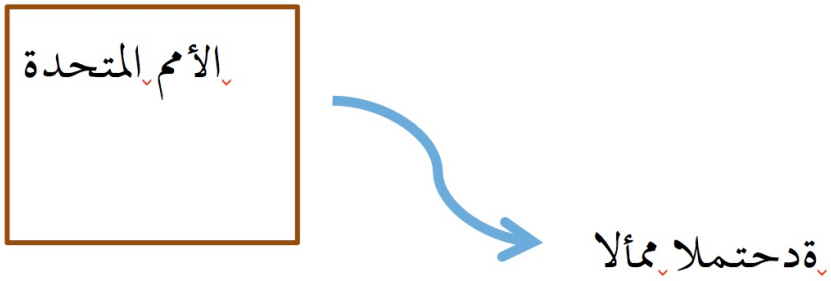




Correct by shifting -350_{hex} code points Unicode Latin range

Bidirectional reversal

Bidirectional reversal



Extraction of RTL text from PDF

- PDF is a display format
- Extracted text is reversed, character-by-character

al‘mm almthdh >> hdhtmla mm’la

Noticeable because the teh-marbuta character [ة] should not begin a word

Sentence-internal fragments

Last year I showed these two slides so that demonstrate that the arctic ice cap, *which for most of the last three million years has been the size of the lower 48 states,* has shrunk by 40 percent.

L'année dernière, je vous ai présenté ces deux diapositives qui montraient que la calotte glacière arctique, *qui pendant ces 3 derniers millions d'année avait la taille des Etats-Unis sans l'Alaska,* ***qui pendant ces 3 derniers millions d'année avait la taille des Etats-Unis sans l'Alaska,*** avait diminué de 40%.

Sentence-internal fragments

+1.53 BLEU on IWSLT'14

Sentence-internal fragments

It's the difference between divergent **thinkingand** convergent thinking. You have to separate

---

the two so that you can diverge your **thoughtsand** come up with this great collection of

---

ideas, and then once you have this great **collectionof** ideas, you focus on the convergent thinking.

Chunking errors in the QED Corpus. Resolved using greedy search for two valid words.

Harvested MT

Weeds of human origin

Under-achieving translation

Errors in human translations:

- Transliteration instead of translation
  Awards and Reviews $\rightarrow$ Награды и Ревью /nagradɨ i rev'ju/
- Codeswitching
  English: In the top ten, India comes in the last
  Urdu: "certification", "top ten", and "number"

Over-achieving translation

Explicitation in human translations:

- Explanations added
- Acronyms expanded
- Visual aids described in target side of TED talks
- Disfluencies removed from translation of speech

Mixed alphabets

она сейчас

Mixed alphabets

Word	Latin (L) or Cyrillic (C)	Meaning
она	LCL	she
сейчас	LCCCCC	now
MP3-плеер	LLL-CCCCC	MP3-player
MP3плеер	LLLCCCCC	MP3player
амазон.com	CCCCCC.LLL	amazon.com
іпациент	LCCCCCCC	iPatient

Russian word она is encoded such that the Latin characters *o* and *a* are used instead of the more appropriate (but visually indistinguishable) Cyrillic equivalents.

Mixed alphabets

Urdu:

- U+002D - Latin Hyphen
- U+06D4 - Urdu Full Stop

French:

- U+00A8 ¨ Latin Diaeresis
- U+0022 " Latin Quotation Mark

Russian:

- U+0431 б Cyrillic Small Letter Be
- U+0036 6 Latin Digit Six

English:

- U+006F o Latin Small Letter O
- U+00B0 ° Latin Degree Sign

Correcting such errors typically requires human intervention

Mixed morphology

Urdu variation in marking plurals on borrowed words

وں lydr + wn/	/	لی‌ڈروں
English plural s	lydr + z/ /	لی‌ڈرز

Mixed use of Urdu and English plural markings

Mixed morphology

Russian variation in marking possession on foreign names:

песню Уитни Хьюстон	song Whitney Huston
закон Артура Кларка	law Arthur+GEN Clark+GEN
Книга Эл Гора	Book Al Gore+GEN

Mixed use of null and Russian plural markings

Mixed morphology

Solutions:

- Stem OOVs before transliterating
- Replace inflected OOV with in-vocabulary variant

Transliteration of Names and Borrowings

Муаммар Каддафи

Transliteration of Names and Borrowings

Муаммар Каддафи

Rule-based transliteration recovery of NE

- Transform target language pronunciation dictionary entries; map to English entries
- Pre-translate known NEs from existing list
- Third-language mappings

Third-language mappings

(**a**) 翟志刚
(**b**) Чжай Чжиган
(**c**) Chzhay Chzhigan
(**d**) Zhai Zhigang

Chinese name (**a**), with transliterations into Cyrillic (**b**) and Latin using normal Cyrillic-to-Latin transliteration (**c**) and reverse Palladius transliteration (**d**). The output in (**d**) is correct.

Weeds among the wheat