Quality Assessment of the Reuters Vol 2 Multilingual Corpus

Robin Eriksson • University of Turku

Turun yliopisto
University of Turku

The Problem

Like every other deliverable, corpora are produced under limited timetable and budget.

Frequently, quality requirements are implicit, informal, and / or nonexistent.

We need better metrics, and tools for assessing and – ideally – improving corpora.

The Corpus

The Reuters Vol. 2 Multilingual Corpus contains newsprint from 1996-1997 from 13 local Reuters offices.

Location	*Articles*
Denmark	11,185
France	85,393
(West) Germany	116,212
Italy	28,406
Japan	65,499
Latin America *	79,818
The Netherlands	17,940
Norway	9,409
Portugal	8,841
Russia	17,487
Spain *	18,655
Sweden	15,732
Taiwan	28,964

With two of the offices (*) publishing in Spanish, but with the (unadvertised) inclusion of English, the corpus contains material in 13 languages.

This multilingual corpus contains material from the same timeframe as the monolingual English Vol 1 corpus.

Each newswire article is exported as a separate XML file with metadata about publication date, categories, etc.

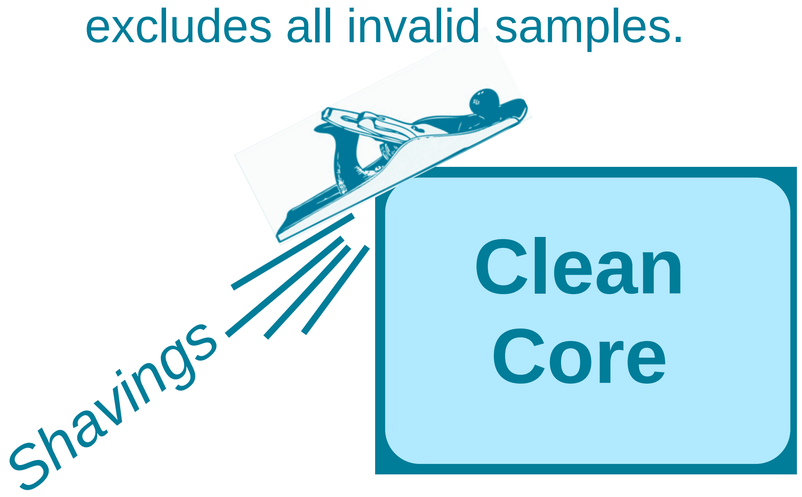
The Methodology

We specify a set of constraints which the samples in the corpus must obey. Violations are excluded or at least investigated.

- *Hard* constraints must never be violated.
- *Soft* constraints may be within normal variation.

By iteratively applying constraints, we shave off undesired artefacts from the clean core of the corpus.

Ultimately, we hope to have an ideal catalog of constraints which excludes all invalid samples.



The Constraints

For this initial case study, a fairly simple and straightforward set of demo constraints was investigated.

- ***No duplicates.***
  In fact, lots of identical samples.
- ***Correct and consistent markup.***
  Are samples represented so that we can understand and manipulate them properly?
- ***Within entropy range.***
  Text in a particular language fits the expected byte entropy distribution. Aberrations may be corrupted or in the wrong language.
- ***Consistent category tagging.***
  Investigation of duplicates and near-duplicates suggests 1% error rate. This is probably too low.

  Also, language only bulk tagged.

The Results

```
<?xml version="1.0" encoding="UTF-8"?>

<newsitem date="1996-08-20" id="root" itemid="0" xml:lang="ru">
<title></title>
<headline> УМВБ - карбованец не изменился в Киеве 176.100 за доллар. </headline>
<byline></byline>
<dateline></dateline>
<text>
<p> МОСКВА, 20 авг (Рейтер) - Курс карбованца к доллару не изменился на Украинской
Межбанковской Валютной Бирже (УМВБ) - 176.100. </p>
<p> 20 авг      19 авг  Курс доллара      176.100     176.100  Объем торгов  ($ млн)
4,33   2,83  Начальное предложение    4,44   3,01  Начальный спрос ($ млн)  4,33   2,83
</p>
<p> -Московское бюро (095) 941 8520. </p>
<p> (c) Reuters Limited 1996. </p></text>
<copyright>(c) 1996 Reuters Limited</copyright>
<metadata>
<codes class="bip:countries:1.0">
<code code="DEVGCOZ">
<editdetail action="confirmed" attribution="Reuters BIP Coding Group" date="1996-08-
20"></editdetail></code>
<code code="EEURZ">
<editdetail action="confirmed" attribution="Reuters BIP Coding Group" date="1996-08-
20"></editdetail></code>
<code code="EURZ">
<editdetail action="confirmed" attribution="Reuters BIP Coding Group" date="1996-08-
20"></editdetail></code>
<code code="UKRN">
<editdetail action="confirmed" attribution="Reuters BIP Coding Group" date="1996-08-
20"></editdetail></code>
<code code="USSRZ">
<editdetail action="confirmed" attribution="Reuters BIP Coding Group" date="1996-08-
20"></editdetail></code></codes>
<codes class="bip:topics:1.0">
<code code="M13">
<editdetail action="confirmed" attribution="Reuters BIP Coding Group" date="1996-08-
20"></editdetail></code>
<code code="M132">
<editdetail action="confirmed" attribution="Reuters BIP Coding Group" date="1996-08-
20"></editdetail></code>
<code code="MCAT">
<editdetail action="confirmed" attribution="Reuters BIP Coding Group" date="1996-08-
20"></editdetail></code></codes>
<dc element="dc.publisher" value="Reuters Holdings Plc"></dc>
<dc element="dc.datepublished" value="1996-08-20"></dc>
<dc element="dc.source" value="Рейтер - Новости на Русском Языке"></dc>
<dc element="dc.creator.location" value=""></dc>
<dc element="dc.creator.location.country.name" value=""></dc></metadata></newsitem>
```

- XML schema problem: Unused elements.
- XML usage: Boilerplate and metadata in sample text.
- XML schema problem: Tags are free-form, uncoordinated.
- Ill-defined space semantics.

Entropy investigations unearthed a number of corruption problems.

```
xml/13982.xml:<p> Из-за отсутствия соглашения с МВФ, Украина с начала года фактически не
получала внешней помощи для покрытия дефицита, несмотря на предварительные договоренности
с международк nikdo z pacientш dal~X9Aьн takovь ручnrad nenahlбsil. </p>
xml/13982.xml:<p> Podle vьzvy MLK by od 1. иervence 1997 mml ka~X9Ehdь obиan u
ambulantnнho lйkaче zaplatit plnou cenu vьkonu podle novйho Seznamu vьkonш ручi hodnotm
bodu ve vь~X9Aьi jednй koruny. Platit za o~X9Aьetчenн si podle MLK budou pacienti do tй
doby, ne~X9Eh bude stanovena reбlnб cena hodnoty bodu. Dohoda zбvisн na dal~X9Aьnch
cenovьch jednбnнch mezi poskytovateli zdravotnн pйиe a zdravotnнmi poji~X9Aь~X9Dkovnami.
</p>
xml/13982.xml:<p> ~XA0 CTAB2923.00b0102070000010300084E7FC05000326100407CD07AB175050000E204
000000000000 00001B0100000000000000000000. </p>
xml/9521.xml:<p> Nach Aussagen Antinoris hat das Staatsunternehmen Air France die erste
Sanierungsphase gerade abgeschlossen. Air France habe seine 37 Milliarden Franc Schulden
durch Finanzspritzen des franzxsischen Staates von insgesamt 20 Milliarden Franc und dem
Verkauf von Unternehmensteilen auf 13 Milliarden Franc reduziert. Zudem habe das
Unternehmen Gesellschaft den Rьckstand zur Lufthansa in den letzten drei Jahren durch den
Abbau von 5000 Stellen, die Senkung der Stьckkosten um 15 Prozent und eine
ujednanй~X83ґmezi~X83ґOrganizacн~X83ґspojenьch~X83ґnбrodш~X83ґa~X83ґIrбk em.~X83ґZ~X83ґBag
dбdu~X83ґto~X83ґdnes~X83ґoznбmily~X83ґzdroje~X83ґz~X83ґOSN.~X80Ъ~X81ґ~X8
3ґ~X83ґ~X83ґ~X83ґ~X83ґBlн~X9Ehe~X83ґnespecifikovanб~X83ґlon~X83ґpodle~X8 3ґjejich
```

```
xml/399940.xml:<p> Weiter im Aufwind befand sich das britische Pfund, das auf 2,4764/4796
(2,4670/4701) sfr kletterte. Spekulationen auf einen
ZinsanstieIO05X06DE07FK040515312GB16R24PL 27B30DEUTSCHLAND/BUNDESWEHR11Bundeswehr-Soldaten
stellten Hinrichtungen nach~17 Bonn (Reuter) - Soldaten der Bundeswehr haben am Rande
einer Ausbildung für den Bosnien-Einsatz Bürgerkriegsszenen wie Vergewaltigungen und
xml/199097.xml:<p> Les députés UDF ont également décidé de constituer un groupe de travail
qui présentera des propositions en matièE E COR MGPQ02RES
ES03REP.MC05T06EN07MD120023612(ES01)16R24MIS30Repsol posts 5.6 pct drop in net profit =217
Net cash flow 183.93 vs 189.00 Earnings per share (pesetas) 289.92 vs 307.16 Cash flow per
share (pesetas) 613.09 vs 629.99 Shareholders&apos; funds 717.82 vs 658.04 Provisions
103.53 vs 95.82 Debt (medium/long-term) 287.89 vs 221.08 Debt (short-term) 163.67 vs
92.73. </p>
xml/371118.xml:<p> 5F8BF0050001100F0E04CD07DA14D010000E2040000000 </p>
xml/15447.xml:<p> Получить комментарий непосредственно в РАО Газпром пока не
уда08309RX02XREF05X06RU07MB290084816R24RUS30Идея о ... XREF17 </p>
xml/336399.xml:<p> LUDWIGSBURG - Wьstenrot Holding GmbH, PK mit erstem Überblick über da
DE ECI05X06DA07SH160006424MIS27RT30buba/inflation11Bundesbank sees inflation staying below
two percent17 FRANKFURT, Jan 16 (Reuter) - The Bundesbank said on Thursday that it
expected German inflation to remain below two percent for the next two years, with high
```

Near-duplicate investigation showed large number of similar samples, but without context, we cannot tell whether they are erroneous or just regular newsroom republication. More investigations needed!

Poster presented at the LREC 2016 in Portorož, Slovenia May 23-28, 2016.

On-line repository: http://github.com/rcv2/