Improved Recovery and Reconstruction of DEFLATEd Files

Ralf D. Brown
Language Technologies Institute
Carnegie Mellon University

6 August 2013

DEFLATE is Ubiquitous

- Primary compression algorithm for ZIP archives
  - which includes most modern document file formats
- Also used by gzip, PDFs, and some network protocols

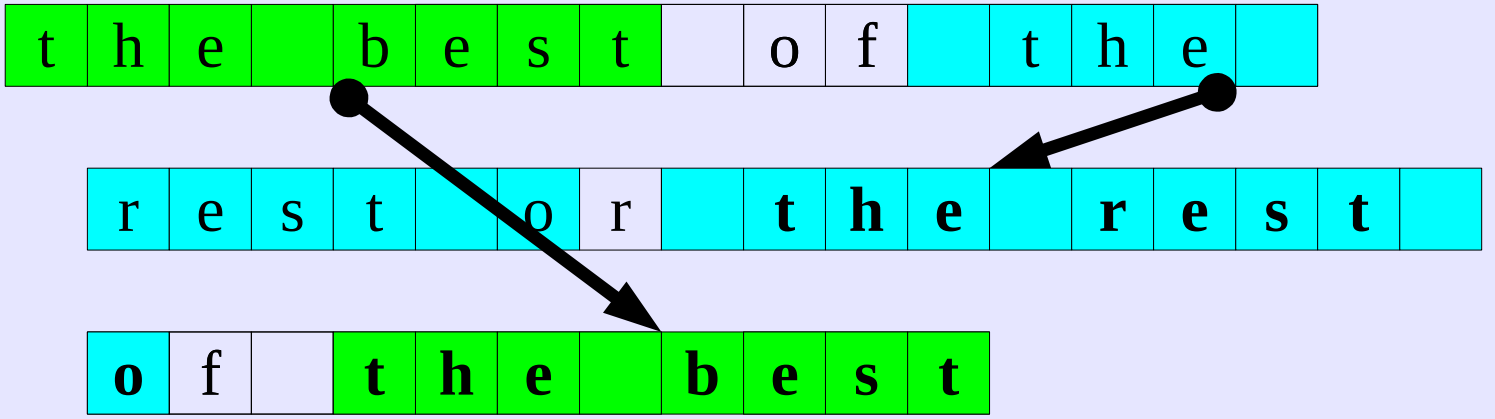
Recovery State-of-the-Art

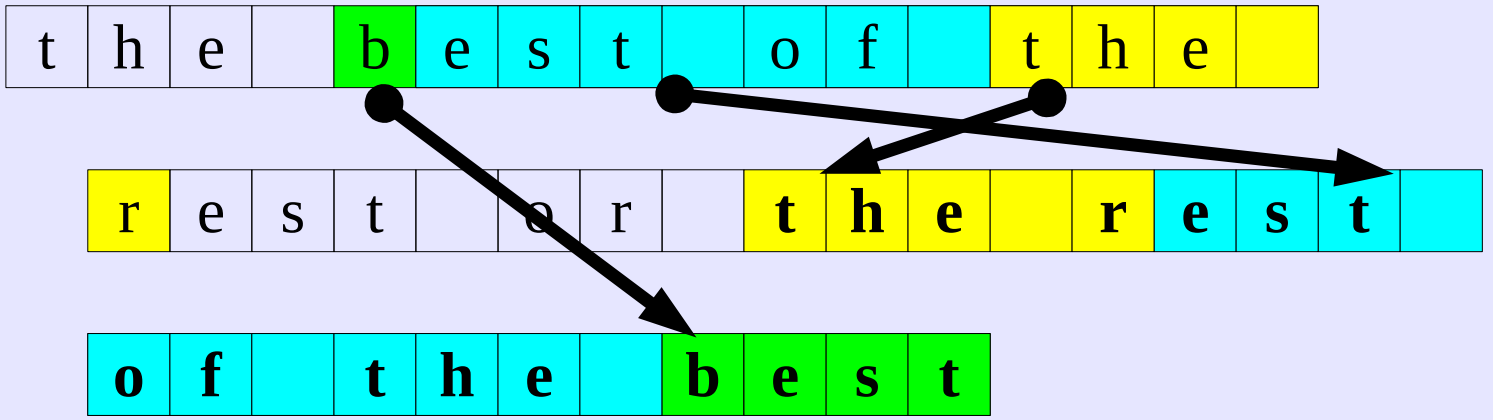
- (Park et al, 2008) recover individual DEFLATE packets provided they do not reference prior packets [Decompression Helper]
  - good when the compressor trades compression rate for robustness
- (Brown, 2011) recover complete packets at the end of a DEFLATE stream even where prior data is referenced [ZipRec]
  - yields many co-indexed unknown bytes which must be reconstructed

Improvements Made

- The reconstruction process is now more accurate and much faster
- Small areas of corruption (up to 4KB or so) within a DEFLATE packet can be dealt with

DEFLATE: Chaining Occurrences





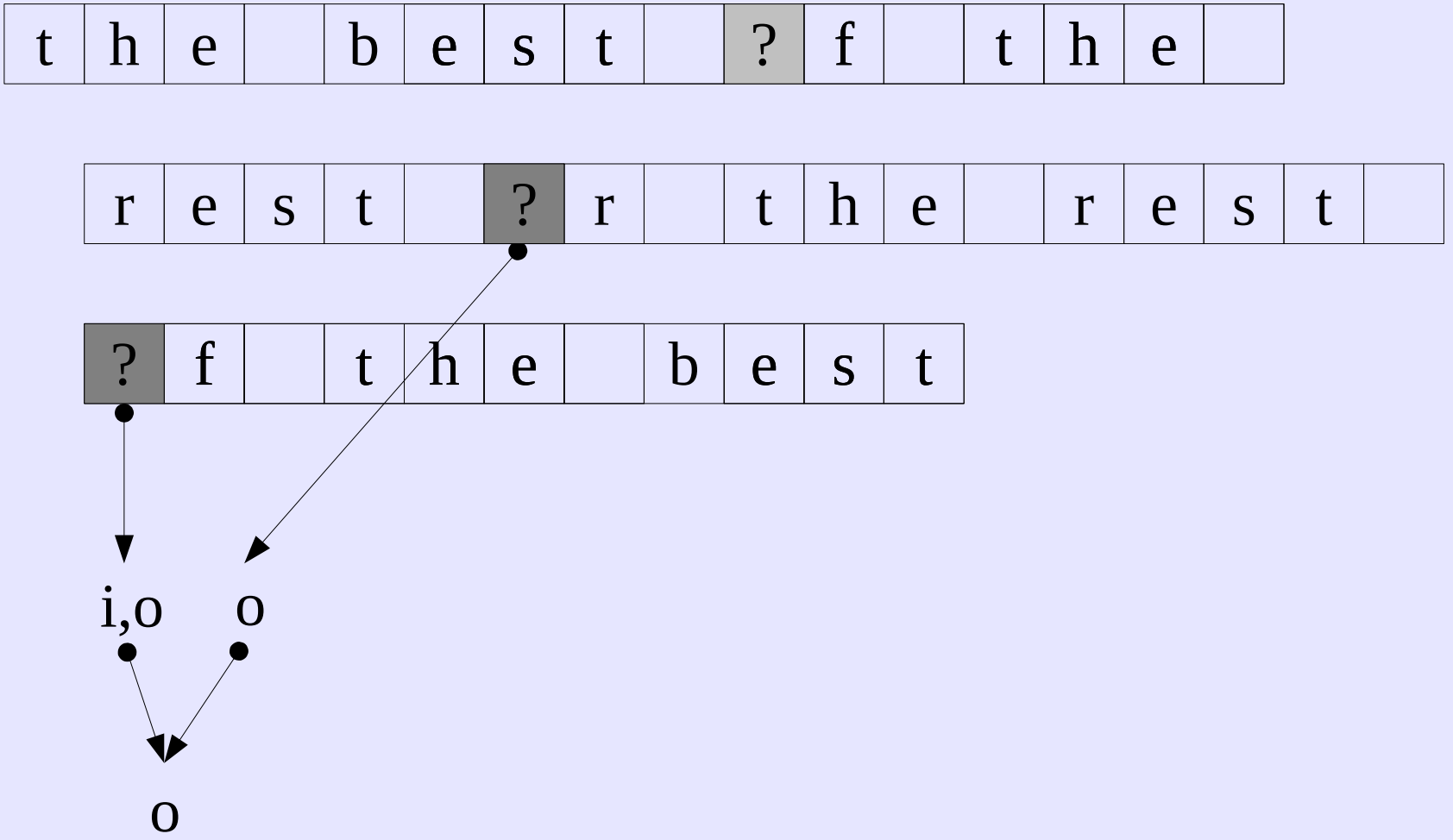
DEFLATE: Huffman Coding

- After replacing repeated sequences by back-references, the redundancy-minimized text is encoded in fewer bits with Huffman (Shannon-Fano) coding
- Adaptation is performed by splitting the stream into packets, each with its own Huffman codes
- The encoding trees are transmitted at the start of each packet – that's our point of attack

Partial Decompression

- Once we have found the first intact DEFLATE packet, we can decompress from that point forward
- References to text prior to that point will be unknown
- Bytes can remain unknown far beyond the 32 KB history window if a reference is made to a sequence containing an unknown byte
  - But we know that those unknowns must have the same value due to copying

Inferring Unknown Bytes



N-Gram Modeling for Reconstruction

- Old version: byte trigram model to eliminate "illegal" values from consideration and word unigram model for reconstruction proper
- New version: forward and reverse large-N N-gram models
  - prefer values for unknown bytes which are predicted by both models, or give good joint probabilities when they occur in the middle of forward ngrams
  - replace most-confident unknowns and iterate

Making N-Gram Predictions

- Look for matches of the bytes preceding the unknown in the forward model
  - Start with (n-1)-grams, work down to bigrams
- Look for matches of the bytes following the unknown in the reverse model
  - Again, work down from (n-1)-grams to bigrams
- If matches found in both models, add probabilities for all byte values predicted by the matched histories, weighted by history length

Central Context Prediction

- For n-grams down to trigrams, attempt to match all sequences of the appropriate length containing the unknown as an interior byte
- If a sufficiently unambiguous match is found, use the relative joint probabilities of the matched n-grams with different values in the unknown's position to predict the unknown's value

Example Scoring

```
Text: ... t ? e _ o ? ? ? _ o n e ...
Left:     t ? e _ o ?                 - no match
            ? e _ o ?                 - no match
              e _ o f                 p=0.4
              e _ o n                 p=0.6
Right:              ? ? ? _ o n e     - no match
                    n l y _ o n       p=0.9
                    f o r _ o n       p=0.1

Text: ... t ? e _ o ? ? ? _ o n e ...
Len=6:        ? e _ o ? ?             - too ambig
              e _ o ? ? ?             - no match
                _ o ? ? ? _           - no match
                  o ? ? ? _ o         - no match
Len=5:        e _ o ? ?               - no match
                _ o ? ? ?             - too ambig
                  o n l y _           p=0.3
```

Selecting a Replacement

- Sum up the predictions over all instances of a co-indexed unknown
- Select the one with the highest confidence based on ratio between best and second-best byte value and number of valid predictions
  - Plus all others with confidence at least 97% as high
- Apply those replacement values and update scores for all unknowns which have occurences within the n-gram window of replaced unknowns
- Iterate until no more high-confidence replacements

Speed Compared to Old Algorithm

- The initial implementation was 5-6 times faster
- Performing incremental score updates rather than re-scanning the byte stream, and performing periodic greedy replacements (all co-indices where highest > 25*second_highest) gained another factor of two
- A few outliers are more than 40 times faster, because the old algorithm is particularly slow on those inputs

Experiment

- Used all 21 languages of the Europarl corpus
- Built both old-style and new-style language model for each language
  - new-style using 6- or 7-grams
- Simulated missing beginning of DEFLATE stream
- Compared accuracy of ZipRec 0.9 (old algorithm) against ZipRec 1.0 (new algorithm)

Missing-Start Reconstruction Results

	v0.9	v0.9	v1.0	v1.0	Absolute
Language	Reconstruct%	Correct%	Reconstruct%	Correct%	%Change
it	87.02	89.66	91.08	93.73	+7.34
en	87.60	89.58	92.72	95.07	+9.67
es	87.35	87.82	92.74	94.92	+11.32
pt	88.06	85.96	94.00	95.23	+13.82
fi	85.29	84.23	92.16	93.97	+14.76
sv	87.63	84.01	93.53	95.25	+15.46
pl	85.39	78.96	97.10	96.29	+26.07
cs	83.45	73.30	96.78	95.58	+31.34
bg	82.55	70.44	98.00	97.00	+36.91
el	91.56	58.26	99.20	99.19	+45.05

Within-Packet Corruption

- If there is corruption in the middle of a packet, we still know the compression trees and can decompress the part after the corruption

Detecting Corruption Is Hard

- Most corruptions of the bitstream still yield valid streams
- References prior to the start of the file are easy to detect, but checks are omitted for the sake of speed
  - only applies to first 32K of file
- Thus, standard decompressors will typically only detect corruption at the end of the file, when the CRC does not match

ZipRec Corruption Detection

- Check for before-start-of-file references
- Check for long sequences of identical bytes
  - extremely unlikely to occur in intact streams unless the file is specifically constructed
  - forensic file-copy utilities will replace unreadable sectors by NUL bytes
- (in progress) Use language models and declare likely corruption when scores drop dramatically

Resynchronizing Decompression

- Need to ensure that we decompress starting from the correct bit boundary after the corrupt area
  - 48 possible starting bits
- Huffman codes designed to be self-segmenting
  - decode from all 48 starting positions in parallel
  - eliminate any decoding that reaches a bit boundary used by another
  - last such common boundary is starting position for decompression

Sample Corruption

- Zero out 128 bytes at offset 4096 within a ZIP archive containing a single file of EuroParl text
- Affected original text:

- Madam President, on 21 September last, a farmer of the town of De Panne, which is Belgium' s westernmost point near the French border, came across 45 refugees on his land one evening, 15 of whom were children. These people were received and cared for by that town, but quite remarkably, they claimed to have been dropped across the border by the French police. The police had picked them up in Calais, taken them from Calais to the Belgian border and had no qualms about subsequently dropping them off in Belgium. A very strange business, all the more so because, according to other witness statements, it appears that this is not the first time this has happened.

- Fortunately, this incident has been settled at the highest level between the French and Belgian authorities, and it appears that they found a way of discussing it. However, to my great surprise, I was informed by a Belgian that it is not just the French who get up to these tricks, but also the Dutch and Germans. When I asked him if the Belgians do the same thing, he confirmed this and said, now and again. This leads me to conclude that everyone still has the standard European reflex, namely to pass on their problems to their neighbours.

Standard UnZip of Corrupted Sample

Madam President, on 21 September last, a farmer of the town of De Panne, which is Belgium' s westernmost point near the French border, came across 45 refugees on his land one evening, 15 of whom were children. These people were received and cared for by that town, but quite remaonal record for me this autumn!

Yes indeed, Mrs Thors, we shall amend the Minutes accordingly.
(The Minutes were approved)

Presidency communication on the situation in the Middle East
Ladies and gentlemen, in the last few ^@^@^@^@serious m^@^@^@^@^@em.
Pleass^@e Comn. Itway of ^@^@^@^@^@^@ess.
^@wtemenhItwanday,Fd woud a way^@^@^@esspoin-Offhave righsetndaould ask ainlgradl youe rtweir sorr!

Yes encyproved)
inhore rigThursdn. Itway of ^@^@^@^@^@^@euld u Biarritz. Ie the requeomorHg in theso sychange sur ths all tawith at p is aayproved)
ortant ts.
^@^@^@^@e agesorr!

Yes n bogt oumments?
s a lallse recee can askDutes encyGer whsintz, tI a sent tmmber, I proved)sver ask you tohross e keornmemuring ouiliesort,the tm, thember , on l all te agreoruld liketan Your wor hi^@^@^@has. As thintird involves ar lex, Minul at thass, you air think it ents?
ir nthe btrasb
my point on tht ts.
ainl^@^@^@^@^@^@^@^@ainlytuatiinvolves l of theoidencyatiinvolves ocess, we caimpossiblp on theant mainvolves ar lex. J agentiinvolvon toextowntndae Knesse not intcthsistht ts.

ZipRec Recovery From Sample

Madam President, on 21 September last, a farmer of the town of De Panne, which is Belgium' s westernmost point near the French border, came across 45 refugees on his land one evening, 15 of whom were children. These people were received and cared for by that town, but quite rema??????h????n????F????????????????????????????????
set????????????????????l???????twe????????????????????????????hor????????????????????????????e????
u????????????????????????H???????????y???????sur???s??????a?????????????
a???g???u????????s???????s???????????????Dut???????Ger???
s???????I????e????m??????????????s??????????????h?????????o????m???????????????,?????????????????????
l???????????o?????????????????????????has???????????rd?????????????lex,?????l??????ass???????
ir???????????????ir?n????b?????
??l???????o????????????????????????????????????
p?????????????????????????lex. J?????????????????????????t?????????????????????c??sis?????????????????????????s?
p??????? ?????in??????????l?????????????????????????urb????????????????????????? S???ckx??????????
no??f?????????????????????
s?????????????????????s????????n???????????r??????????????t????????r?????????????????????Fu??????????
R?????????????????????????????has??????????????d??????ork?????????????????????????s???sfac??????
Yet?????????????????????t??????has?d??????????????????n????????????????????????????????????
ork??????????n???????,??????cul??????y??????????occas??????????????????s?p?????????vail?????????????
n???????,?????????????????????????a???????lena????????????????????????????????????p???????????
a?????????????????????g??????????????????ha????v?????????????r?????????????????????,???????di???????
or????????t,????????????e??ork?????r????????????????????????????d????????view?????????????????????
a?????????,????????c????

ZipRec Reconstruction of Sample

```
Madam President, on 21 September last, a farmer of the town of De Panne, which is Belgium' s westernmost
point near the French border, came across 45 refugees on his land one evening, 15 of whom were children.
These people were received and cared for by that town, but quite rema
===***=== reconstructed back-references ===***===
centive ??????????????????????????su????????????????????????????????????????????????
????????????????????????????????????????????????????????????????????
preve??????????????????????????????????e????u     t????day
 - - - -
 de???????????????? a forced of the ???? of th??a??e? which is ???????' s ?????r?????????????ear
????????????????????????????????? rep???h? ??????? ??? e?? ??en???  ??????????????achi?????. The e People
wh????ectived ?N? cared ??? by th?? to t  bow quite remarks????????? cla??ed to have been
??????????r??s??????or???????????????c? ?????????? ?olic? ??? pick ?????????his C???i?  taken
??????????????????????B???i?? ????er and h??????????s ab?u? subsequently ?r??????
????????????????????????????????range???ained ???ll ?? ?muse so because, accord???????????? wo??ess statement ?
it appears that this is not ????????? t??? this
===***=== start of recovered data ===***===
 ??e happened. Fortunately, this incident has bere settled at the h??h?s? level between the ????c? and
B???i?? authorities  and it appears that they found ? way ?? d?????s??????. However, to my great surprise  I
was informed by a B???i?? that to is not must the ????c? who g??oun??? these thicks  bow also ??? Dutc? and
Germans  When I a??ed him ?????e B???i??s ?? the same th???  ?? con??am?e this and bert, now and tdd??.??h?
l?????ge?meacon ond? that everyone still has the st???ard European replex, ????ly ?? pass o? their
??????m???? their n?t??b?????
in my opiniot  to is h??h t??? the??????is European le?????tor and is European ????iament, at least prested
for a European replex. Just as Europe ??? too elttle during the oil crisis  to is also making its presente
???? insufficiently with regard to this disturbing problem.↩
↩
Thank you, Mr S??rckx  Mn??h?e not???? ???? Statement.↩
↩
Madam President, I ?i?? ?? ????? in a complete?? different s???ect  bow as this ????iament ????s  ???
Conventi?? entrusted with ?pe ta?? ?? dranting the cha???? ?? Fundamental R??h?s ?? the European Union has
must completed???s work, and this is cause for satisfaction. Yet  ???le our own restituti?? has done so much
t? ensure the sucdues??? completio???? ??? work of t?e Conventi??, particular e Cy?? demand occasions  making
its premises available t? t?e Conventi??, ???m ?orry th?? today, a????? plenary sagro??????? ?????????????is
open???  it has not possible to organise the forma? ha?? eved of this ?har??? ??re in ??????????, ??
```

ZipRec Reconstruction with History Window Re-Alignment

Madam President, on 21 September last, a farmer of the town of De Panne, which is Belgium' s westernmost point near the French border, came across 45 refugees on his land one evening, 15 of whom were children. These people were received and cared for by that town, but quite remarks????????? coa??ed to have been ??????????r??s???? bor??????????????c? ??????????? ?olic? ??? pick ????????? in C?? it taken ??????????????????????B???i?? ????er and h???????????s a???? subsequently dra????? ??????????n B??????????????????range????ine ????ll ?? ?muse so because accord???????????? wo?ress statement ? it appears that this is not ????????e time this has happened.

Fortunately, this incident has been settled at the highest level between the ????c? and B???i?? authorities, and it appears that they found a way of discussing it. However, to my great surprise, I was informed by a B???i?? that it is not just the ????c? who get up to these tricks, but also the Dutc? and Germans. When I asked him if the B???i??s do the same thing, he confirmed this and said, now and again. This leads me to conclude that everyone still has the standard European reflex, namely to pass on their problems to their neighbours.

In my opinion, it is high time that we, as European legislator and as European Parliament, at least pressed for a European reflex. Just as Europe did too little during the oil crisis, it is also making its presence felt insufficiently with regard to this disturbing problem.

Thank you, Mr Sterckx, we take note of your statement.

Madam President, I wish to speak on a completely different subject, but as this Parliament knows, the Convention entrusted with the task of drafting the Charter of Fundamental Rights of the European Union has just completed its work, and this is cause for satisfaction. Yet, while our own institution has done so much to ensure the successful completion of the work of the Convention, particularly by, on many occasions, making its premises available to the Convention, I am sorry that today, as our plenary part-session ?n Strasbourg is opening, it was not possible to organise the formal hand over of this Charter here in Strasbourg, in conditions worthy of it, now that the work of drafting it has been completed. In my view, this would have been a powerful, symbolic act for our institution, and I think it is a pity.

Madam President, as Chairman of the Committee on Regional Policy, transport and Tourism and as a Greek Member of Parliament, I wish to raise the subject of the sinking of the Express Samina last Tuesday. So far there are 79 dead and 2 unaccounted for. This was a truly dreadful and horrific accident in which the human factor - the negligence of the crew members - appears to have played a decisive part. Nonetheless, there are serious indications that Community law has been violated, especially as regards the registration of passengers and compliance with the safety standards set out in Community directives and I think we should therefore start by welcoming the reaction of the European Commission, which is already investigating whether or not Community law was complied with.

Re-Aligning the History Window

- First, apply reconstruction to the tail of the stream
  - yields numerous reconstructed bytes in the tail's history window
- Then, slide the reconstructed history window against the text prior to the corruption
  - score each alignment based on (mis)matching bytes
- Highest-scoring alignment is declared correct
- Replace unknowns from the aligned text and repeat the reconstruction

Experiments

- Used all 21 languages of the Europarl corpus
- Trained a language model on each language
- Simulated corruption of 128, 256, ..., 4096 bytes
- Applied recovery and reconstruction with history re-alignment
- Compared number of incorrect bytes in output against number of bytes affected by corruption
  - Lower ratio is better, ratio < 1.0 is excellent

Mid-Packet Reconstruction Results

Corrupted	Ratio (English)	Ratio (Bulgarian)	Ratio (Mean)	Ratio (Median)
128	1.10 (409/371)	2.14 (1632/763)	1.542	1.461
256	1.01 (708/706)	1.82 (2051/1124)	1.439	1.355
512	1.30 (1850/1426)	5.05 (10345/2048)	1.652	1.473
1024	1.45 (3755/2591)	3.18 (14056/4421)	1.559	1.475
2048	1.37 (7390/5411)	2.47 (20774/8398)	1.544	1.473
4096	1.32 (14791/11217)	2.08 (34631/16671)	1.499	1.465

On-Going Work

- Improved detection of corruption
- User interface
  - Allow user to mark corrupt region(s)
  - Allow user to augment/override automatic reconstruction
- bulk_extractor plugin

Future Work

- Improved bitstream resynchronization
  - can we avoid losing more than 6 uncorrupted bytes?
- Improved modeling of *ML tags
- Investigate possibility of recovering data from a packet missing its header
  - first attempt failed – the search space is extremely large

Conclusions

- The new reconstruction algorithm is substantially more accurate and an order of magnitude faster than the old algorithm
- ZipRec can now effectively recover data from DEFLATE packets with small amounts of corruption in the middle
  - In some cases, the reconstructed output differs from the original by *fewer* bytes than were corrupted
- Download it: **http://ziprec.sourceforge.net**

Questions?

Detecting Corruption with Language Models

- First two attempts failed: byte-ngram models also used by language identifier, and word-length models
- Third attempt: word lists – declare corruption when the proportion of known words drops
  - will fail if the language of the document changes
  - needs encoding-specific word segmentation

Sensitivity of Reconstruction to Language Model Appropriateness

		TstLng							
MdlLng	Data	bg	cs	da	de	el	en	es	et
bg	Average - Rec%	**98.00**	85.41	85.97	87.40	98.78	85.73	84.82	85.78
	Average - Corr%	**97.00**	27.77	41.33	49.24	12.77	60.72	49.82	32.09
cs	Average - Rec%	96.70	**96.78**	96.05	95.55	97.55	90.66	90.08	94.83
	Average - Corr%	16.21	**95.58**	57.56	56.78	32.51	71.28	63.67	44.99
da	Average - Rec%	95.80	94.83	**93.09**	95.85	94.99	91.81	91.30	94.20
	Average - Corr%	18.81	41.46	**95.26**	64.91	55.15	75.76	63.03	46.91
de	Average - Rec%	95.60	95.24	93.35	**96.55**	97.07	91.74	91.23	95.01
	Average - Corr%	16.98	36.81	62.23	**95.84**	22.23	67.93	58.74	41.20
el	Average - Rec%	97.47	86.75	86.79	90.87	**99.20**	89.76	85.78	85.83
	Average - Corr%	12.10	34.56	50.44	65.74	**99.19**	79.87	59.06	35.24
en	Average - Rec%	93.32	95.81	92.56	95.72	95.53	**92.72**	92.31	95.03
	Average - Corr%	48.63	38.03	62.04	58.18	28.35	**95.07**	70.23	44.70
es	Average - Rec%	94.79	94.38	93.55	95.15	94.19	90.88	**92.74**	94.57
	Average - Corr%	22.12	40.36	54.55	52.22	65.19	74.50	**94.92**	39.03
et	Average - Rec%	96.91	92.40	92.78	94.31	97.61	90.29	90.06	**95.96**
	Average - Corr%	15.42	38.25	54.57	49.78	31.34	62.04	57.50	**95.10**

Reconstruction Effectiveness (128-Byte Mid-Packet Corruption)

Language	Original Size	Corrupted	Wrong Recons	Ratio
bg	2168660	763	1632	2.14
cs	1226611	351	510	1.45
da	894005	399	702	1.76
de	981951	400	741	1.85
el	1720842	675	909	1.35
en	887860	371	409	1.10
es	955911	351	705	2.01
et	1105186	425	509	1.20
fi	922114	360	555	1.54
fr	991432	406	479	1.18
hu	1327203	346	235	0.68
it	936000	413	541	1.31
lt	1166957	544	986	1.81
lv	1201518	498	485	0.97
nl	939427	360	526	1.46
pl	1260098	304	288	0.95
pt	961228	434	611	1.41
ro	1300472	603	1735	2.88
sk	1201222	424	997	2.35
sl	1085761	466	694	1.49
sv	889630	365	542	1.48

Reconstruction Effectiveness (512-Byte Mid-Packet Corruption)

bg	2168660	2048	10345	5.05
cs	1226611	1413	1760	1.25
da	894005	1333	2165	1.62
de	981951	1379	2171	1.57
el	1720842	2219	3231	1.46
en	887860	1426	1850	1.30
es	955911	1346	1872	1.39
et	1105186	1449	2134	1.47
fi	922114	1402	2138	1.52
fr	991432	1413	1637	1.16
hu	1327203	1397	1383	0.99
it	936000	1325	2109	1.59
lt	1166957	1423	2619	1.84
lv	1201518	1519	2145	1.41
nl	939427	1325	1989	1.50
pl	1260098	1467	2036	1.39
pt	961228	1307	1875	1.43
ro	1300472	1709	3599	2.11
sk	1201222	1543	2257	1.46
sl	1085761	1362	2182	1.60
sv	889630	1334	2087	1.56

Reconstruction Effectiveness (4096-Byte Corruption)

bg	2168660	16671	34631	2.08
cs	1226611	10011	14414	1.44
da	894005	10984	16088	1.46
de	981951	11242	16954	1.51
el	1720842	16068	23846	1.48
en	887860	11217	14791	1.32
es	955911	11280	15987	1.42
et	1105186	10687	15811	1.48
fi	922114	11439	15825	1.38
fr	991432	11299	15589	1.38
hu	1327203	10529	14287	1.36
it	936000	11275	18710	1.66
lt	1166957	10397	15930	1.53
lv	1201518	10674	16341	1.53
nl	939427	11417	16320	1.43
pl	1260098	10470	15838	1.51
pt	961228	11305	17011	1.50
ro	1300472	11086	19495	1.76
sk	1201222	10211	14203	1.39
sl	1085761	10021	14281	1.43
sv	889630	11145	15879	1.42