

New Insights in Context Modeling *for Text Compression*

M. Oğuzhan Külekci

kulekci@uekae.tubitak.gov.tr

TÜBİTAK-BİLGEM

National Research Institute of Electronics&Cryptology, Turkey

Özyeğin University
İstanbul, Turkey
September 14th, 2011

Prediction

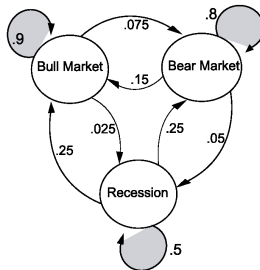


Michel de Nostredame (1503–1566)

Prediction Equipment



Prediction Equipment



Markov Chain

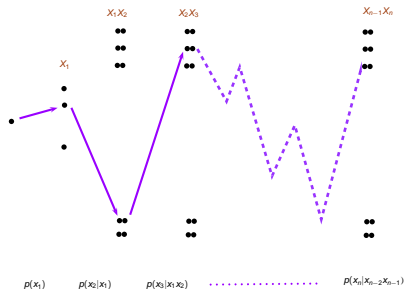
Sequence $\mathcal{X} = X_1 X_2 \dots X_n$, where $X_i \in \Sigma = \{\epsilon_1, \epsilon_2, \dots, \epsilon_\sigma\}$.

Basic Markov Property

$$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \dots, X_{i-1} = x_{i-1}) = P(X_i = x_i | X_{i-1} = x_{i-1})$$

Order- m Markov Chain

$$P(X_i = x_i | X_1 = x_1, X_2 = x_2, \dots, X_{i-1} = x_{i-1}) = P(X_i = x_i | X_{i-1} = x_{i-1}, X_{i-2} = x_{i-2}, \dots, X_{i-m} = x_{i-m})$$



Data Compression

- Main goal: Squeeze data to its information content (*as close as possible*).
- In general, it is a two-phase process:
 - 1 Modeling (dictionary, grammar, statistical, ...)
 - 2 Coding (Huffman coding, arithmetic coding, ...)

Data Compression

- Main goal: Squeeze data to its information content (*as close as possible*).
- In general, it is a two-phase process:
 - 1 **Modeling** (dictionary, grammar, statistical, ...)
 - 2 Coding (Huffman coding, arithmetic coding, ...)
- Improved modeling \Rightarrow Improved compression ratios
- Among the various compression methods, we will focus on context-based *lossless* compression schemes.

PPM

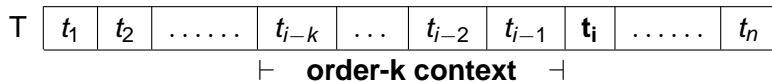
- Prediction by partial matching (CW'84) variants are among the best *lossless* compression methods.
- Main idea: Predict next symbol based on the preceding symbols.



- Compress t_i by encoding the probability $P(t_i | t_{i-1} \dots t_{i-k})$.

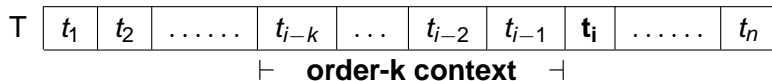
PPM Compression Mechanism

Assume t_i is to be encoded.



PPM Compression Mechanism

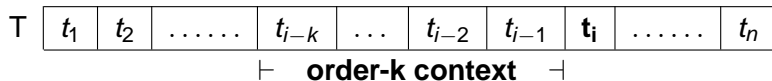
Assume t_i is to be encoded.



- ❶ If $[t_q \dots t_{q+k}] = [t_{i-k} \dots t_i]$ for $\exists q : q < (i - k)$
 then send $P(t_i | [t_{i-k} \dots t_{i-1}])$ to the encoder.

PPM Compression Mechanism

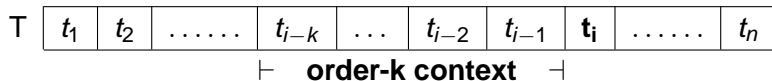
Assume t_i is to be encoded.



- 1 If $[t_q \dots t_{q+k}] = [t_{i-k} \dots t_i]$ for $\exists q : q < (i - k)$
then send $P(t_i | [t_{i-k} \dots t_{i-1}])$ to the encoder.
- 2 Else **emit escape symbol**,
change context (*usually decrementing k by 1*),
if new context is non-empty, go to previous step
else encode t_i by its frequency.

PPM Compression Mechanism

Assume t_i is to be encoded.



- 1 If $[t_q \dots t_{q+k}] = [t_{i-k} \dots t_i]$ for $\exists q : q < (i - k)$
then send $P(t_i | [t_{i-k} \dots t_{i-1}])$ to the encoder.
- 2 Else **emit escape symbol**,
change context (*usually decrementing k by 1*),
if new context is non-empty, go to previous step
else encode t_i by its frequency.
- 3 Continue encoding with the next character, $i = i + 1$.

Difficulties in PPM

Sparseness and zero-frequency problem

- We need to collect $\sum_{i=1}^k \sigma^i$ statistics in an order- k context model.
- In practice, most of them are missing or observed very rare, especially on high orders.
- Long escape sequences when novel symbols are encountered.
- Effective calculation of symbol and escape probabilities has not been totally solved yet.

Difficulties in PPM

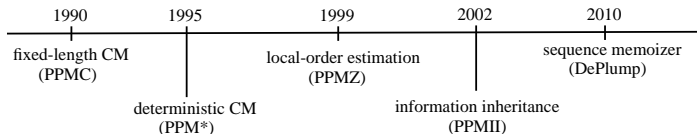
Huge memory requirements due to complicated data structures

- Context trees to store and retrieve statistics.
- Or use tables with hashing for simpler solutions.

Slow speed (*as a consequence*)

- Retrieval and update of statistics take time.
- Change of context at novel symbols takes time.

Improvements in PPM



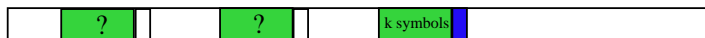
Two main points of research to improve PPM:

- 1 Symbol and escape probability calculations
PPMA, PPMB, PPMC, PPMX, PPMP, PPMII, DePlump

	Symbol Probability	Escape Probability
PPMA	$\frac{c_j}{1+C}$	$\frac{1}{1+C}$
PPMB	$\frac{c_j-1}{C}$	$\frac{\rho}{C}$
PPMC	$\frac{c_j}{C+\rho}$	$\frac{\rho}{C+\rho}$

2 ALTERNATIVE CONTEXT-MODELING TECHNIQUES

Approaches in Context Modeling



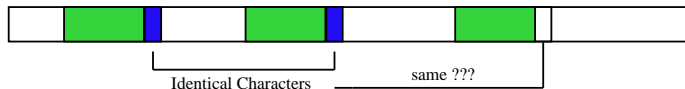
Bounded-length CM

- Maximum context length is predetermined.
- Basic scheme, easy to implement.

Approaches in Context Modeling

Unbounded-length CM

- **Deterministic CM:** Seek for the context that is always followed by the same symbol.



- If no such deterministic context is available, switch to bounded model.
- Heavy computational load, not that much gain.

Variable Length Context Models



f(?.) - characters

- Context length is unbounded (?), but not required to be deterministic.
- Rank context lengths due to some heuristics, select the best value
 - local order estimation** in PPMZ(1999)

$$a \cdot P \cdot \log P + b \cdot E \cdot (\log E - c) + d \cdot (1 - P) \cdot (\log E - c)$$
 - variable context** in PPMVC(2004).
Parameters: d, MinRML, MinLML, and pointers per context
- More complicated, more memory, so much **black-art**

An intuitive example...

mis

An intuitive example...

mis

misanthropic

mis**b**ehave

misc**c**ellany

mis**d**eed

mis**e**rable

misfortune

mis**g**uided

mis**h**ap

mis**i**nform

mis**j**udge

mis**l**ead

mis**m**atch

mis**o**gynist

mis**p**lace

mis**r**epresent

mis**t**ake

mis**u**se

An intuitive example...

mis

misanthropic

mis**b**ehave

misc**c**ellany

mis**d**eed

mis**e**rable

misfortune

mis**g**uided

mis**h**ap

misinform

mis**j**udge

mislead

mis**m**atch

mis**o**gynist

mis**p**lace

misrepresent

mistake

mis**u**se

fir

An intuitive example...

mis

misanthropic

mis**b**ehave

misc**e**llany

mis**d**eed

mis**e**rable

mis**f**ortune

mis**g**uided

mis**h**ap

mis**i**nform

mis**j**udge

mis**l**ead

mis**m**atch

mis**o**gynist

mis**p**lace

mis**r**epresent

mis**t**ake

mis**u**se

fir

fire

firm

first

Information Content of Contexts are Variable

- The information contents carried by different contexts of a symbol are not equal.

... adequate, $P(u|q)$ is possibly good enough.

However;

... obvious, $P(u|o)$???

... obvious, $P(u|io)$??

... obvious, $P(u|vio)$?

- Thus, the context length should be variable!

"WHAT IS THE BEST CONTEXT LENGTH AT A POSITION?"

Correct decision requires enough evidence

- We need **enough amount of information**.

Correct decision requires enough evidence

- We need **enough amount of information**.

Previous definition of *context*

The context of a symbol at a particular position is its **preceding k symbols**.

New definition of *context*

The context of a symbol at a particular position is its **preceding k -bits information**.

Correct decision requires enough evidence

- We need **enough amount of information**.

Previous definition of *context*

The context of a symbol at a particular position is its **preceding k symbols**.

New definition of *context*

The context of a symbol at a particular position is its **preceding k -bits information**.

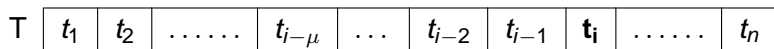
How to compute the preceding k -bits of information for all positions in the text?

Main Idea

- The amount of information can be calculated by compression !
- We can measure the previous information content of t_i by compressing its preceding characters $t_{i-1}t_{i-2} \dots$
- Something like **compression-by-compression** ?

Compressed Context Modeling

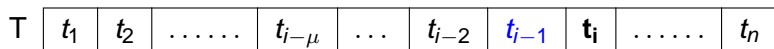
Compressed context of t_i is its preceding \mathcal{I} -bit information.



Bit-array $B = b_1 b_2 \dots b_\ell = \text{compress}(S = s_1 s_2 \dots s_\omega)$, S is character array.

Compressed Context Modeling

Compressed context of t_i is its preceding \mathcal{I} -bit information.

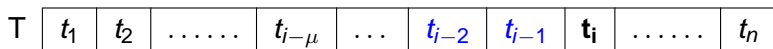


- $B = \text{compress}(t_{i-1})$, while $|B| < \mathcal{I}$, continue ...

Bit-array $B = b_1 b_2 \dots b_\ell = \text{compress}(S = s_1 s_2 \dots s_\omega)$, S is character array.

Compressed Context Modeling

Compressed context of t_i is its preceding \mathcal{I} -bit information.



- $B = \text{compress}(t_{i-1})$, while $|B| < \mathcal{I}$, continue ...
- $B = \text{compress}(t_{i-1}t_{i-2})$, while $|B| < \mathcal{I}$, continue ...

.....

Bit-array $B = b_1b_2 \dots b_\ell = \text{compress}(S = s_1s_2 \dots s_\omega)$, S is character array.

Compressed Context Modeling

Compressed context of t_i is its preceding \mathcal{I} -bit information.



- $B = \text{compress}(t_{i-1})$, while $|B| < \mathcal{I}$, continue ...
- $B = \text{compress}(t_{i-1}t_{i-2})$, while $|B| < \mathcal{I}$, continue ...
-
- $B = \text{compress}(t_{i-1}t_{i-2} \dots t_{i-\mu})$, stop when $|B| \geq \mathcal{I}$.

Bit-array $B = b_1b_2 \dots b_\ell = \text{compress}(S = s_1s_2 \dots s_\omega)$, S is character array.

Compressed Context Modeling

Compressed context of t_i is its preceding \mathcal{I} -bit information.



- $B = \text{compress}(t_{i-1})$, while $|B| < \mathcal{I}$, continue ...
- $B = \text{compress}(t_{i-1}t_{i-2})$, while $|B| < \mathcal{I}$, continue ...
-
- $B = \text{compress}(t_{i-1}t_{i-2} \dots t_{i-\mu})$, stop when $|B| \geq \mathcal{I}$.
- The context is the first \mathcal{I} bits of $B = \text{compress}(t_{i-1}t_{i-2} \dots t_{i-\mu})$.

Bit-array $B = b_1b_2 \dots b_\ell = \text{compress}(S = s_1s_2 \dots s_\omega)$, S is character array.

Measuring the Performance

- Files in *large* Calgary corpus are modeled with compressed context modeling.
- Average empirical entropies are measured as **bits-per-character** (average log-loss).

$$\sum_{\forall \mathcal{C}} p(\mathcal{C}) \cdot \sum_{\forall \epsilon \in \mathcal{C}} p^{\mathcal{C}}(\epsilon) \cdot \log p^{\mathcal{C}}(\epsilon)$$

- The compression function *comp()* used in CCM is a simple 0-order static Huffman coding.
- Comparisons achieved against classical definition of context for lengths from 1 to 10.

Performance Comparison on Equal-Length Contexts

Order- k **byte** context versus order- $8 \cdot k$ **bit** compressed context.

		Empirical Entropy in bits-per-character	
	File Name	CM-1	CCM-8
1	bib	3.36	2.85
2	book1	3.58	3.10
3	book2	3.74	3.23
4	geo	4.26	4.16
5	news	4.09	3.63
6	obj1	3.46	3.13
7	obj2	3.87	3.64
8	paper1	3.64	3.01
9	paper2	3.52	2.91
10	paper3	3.55	2.99
11	paper4	3.47	2.67
12	paper5	3.52	2.71
13	paper6	3.61	2.91
14	pic	0.82	0.75
15	progc	3.60	2.86
16	progl	3.21	2.53
17	progp	3.18	2.45
18	trans	3.35	2.78

Performance Comparison on Equal-Length Contexts

Order- k **byte** context versus order-8 · k **bit** compressed context.

		Empirical Entropy in bits-per-character			
	File Name	CM-1	CCM-8	CM-2	CCM-16
1	bib	3.36	2.85	2.30	1.28
2	book1	3.58	3.10	2.81	1.95
3	book2	3.74	3.23	2.73	1.68
4	geo	4.26	4.16	3.65	2.24
5	news	4.09	3.63	2.92	1.72
6	obj1	3.46	3.13	1.75	0.68
7	obj2	3.87	3.64	2.52	1.68
8	paper1	3.64	3.01	2.33	1.16
9	paper2	3.52	2.91	2.51	1.32
10	paper3	3.55	2.99	2.55	1.21
11	paper4	3.47	2.67	2.20	0.82
12	paper5	3.52	2.71	2.04	0.77
13	paper6	3.61	2.91	2.25	1.07
14	pic	0.82	0.75	0.81	0.44
15	progc	3.60	2.86	2.13	1.05
16	progl	3.21	2.53	2.04	1.02
17	progp	3.18	2.45	1.75	0.94
18	trans	3.35	2.78	1.97	1.01

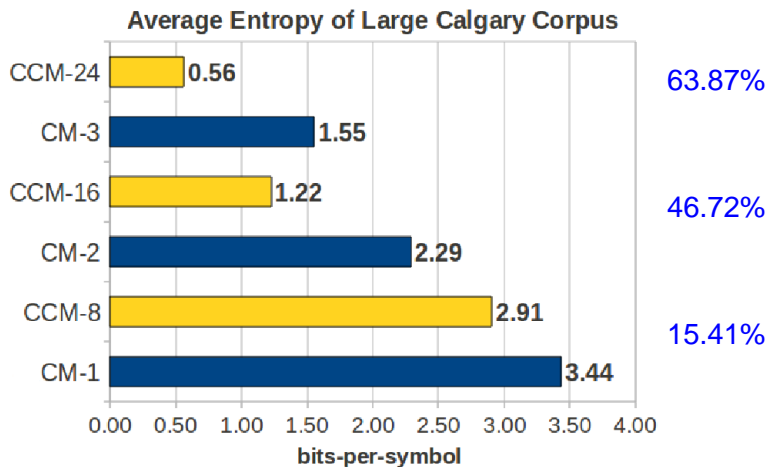
Performance Comparison on Equal-Length Contexts

Order- k **byte** context versus order-8 · k **bit** compressed context.

		Empirical Entropy in bits-per-character					
File Name		CM-1	CCM-8	CM-2	CCM-16	CM-3	CCM-24
1	bib	3.36	2.85	2.30	1.28	1.38	0.67
2	book1	3.58	3.10	2.81	1.95	2.19	1.21
3	book2	3.74	3.23	2.73	1.68	1.88	0.99
4	geo	4.26	4.16	3.65	2.24	3.61	0.20
5	news	4.09	3.63	2.92	1.72	1.81	0.75
6	obj1	3.46	3.13	1.75	0.68	1.40	0.22
7	obj2	3.87	3.64	2.52	1.68	2.20	0.81
8	paper1	3.64	3.01	2.33	1.16	1.40	0.56
9	paper2	3.52	2.91	2.51	1.32	1.68	0.66
10	paper3	3.55	2.99	2.55	1.21	1.60	0.54
11	paper4	3.47	2.67	2.20	0.82	1.18	0.32
12	paper5	3.52	2.71	2.04	0.77	1.05	0.33
13	paper6	3.61	2.91	2.25	1.07	1.29	0.53
14	pic	0.82	0.75	0.81	0.44	0.71	0.20
15	progc	3.60	2.86	2.13	1.05	1.21	0.53
16	progl	3.21	2.53	2.04	1.02	1.23	0.55
17	progp	3.18	2.45	1.75	0.94	1.01	0.54
18	trans	3.35	2.78	1.97	1.01	1.12	0.53

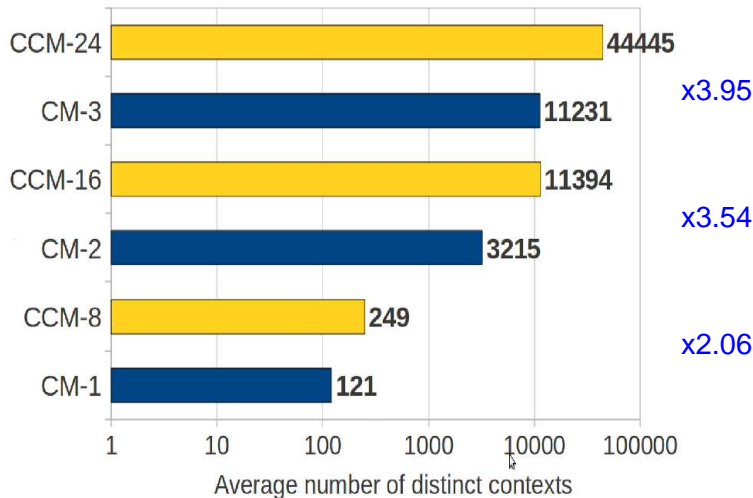
Performance Comparison on Equal-Length Contexts

Order- k **byte** context versus order- $8 \cdot k$ **bit** compressed context.



However, ...

The number of distinct contexts in compressed context is much higher.



Performance of space-restricted compressed context

Choose largest α such that the number of distinct contexts in order $-\alpha$ compressed context is less than that of in order $-\beta$ context.

Context Length (bits) (bytes)		# of distinct contexts		Avg. Empirical Entropy in bpc		Percentage of Gain in Space	Percentage of Gain in Entropy
CCM	CM	CCM	CM	CCM	CM		
6	1	64	95	3.61	3.64	32.63	0.9
11	2	1477	1556	2.13	2.33	5.08	8.62
15	3	4769	6155	1.47	1.40	22.52	-4.55
18	4	11324	12841	0.94	0.90	11.81	-4.84
22	5	19412	19841	0.61	0.63	2.16	2.09
24	6	20920	26074	0.56	0.44	19.77	-27.74

Avg.: 15.66 Avg.: -4.25

The result of space-restricted comparison on file *paper1*.

Performance of space–restricted compressed context

Context Length (bits) (bytes)		# of distinct contexts		Avg. Empirical Entropy in bpc		Percentage of Gain in Space	Percentage of Gain in Entropy
CCM	CM	CCM	CM	CCM	CM		
6	1	64	82	3.49	3.58	21.95	2.46
11	2	1775	1826	2.59	2.81	2.79	7.83
15	3	13187	13296	2.06	2.19	0.82	6.08
18	4	39145	49956	1.74	1.73	21.64	-0.32
22	5	107729	124119	1.38	1.37	13.21	-0.41
24	6	154136	227992	1.21	1.05	32.39	-15.99

Space-restricted comparison on file *book1*

Avg.: 15.47

Avg.: -0.06

Context Length (bits) (bytes)		# of distinct contexts		Avg. Empirical Entropy in bpc		Percentage of Gain in Space	Percentage of Gain in Entropy
CCM	CM	CCM	CM	CCM	CM		
6	1	249	256	3.13	3.46	2.73	9.49
13	2	3802	4766	1.26	1.75	20.23	27.85
16	3	7686	8246	0.68	2.40	6.79	51.67
18	4	9551	9793	0.46	1.25	2.47	62.74
19	5	10252	10614	0.39	1.24	3.41	68.06
20	6	10832	11147	0.34	1.16	2.83	70.52

Space-restricted comparison on file *obj1*

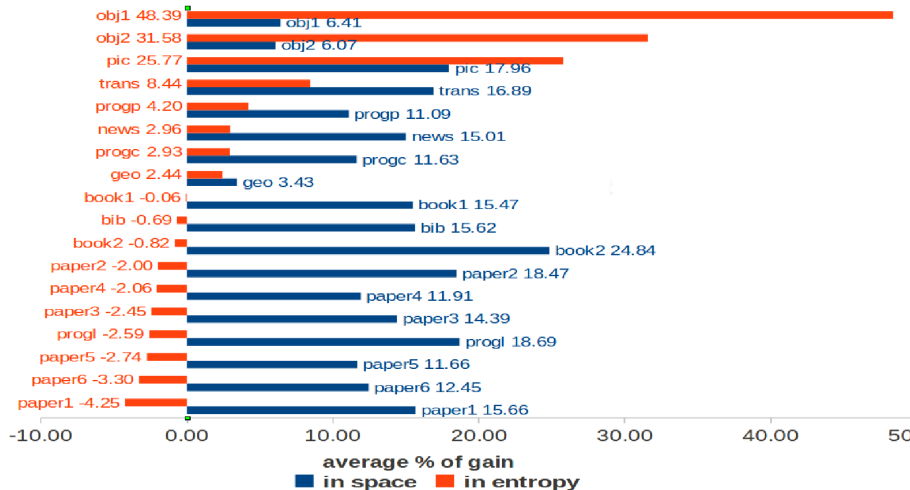
Avg.: 6.41

Avg.: 48.39

Overall Space–Restricted Comparisons

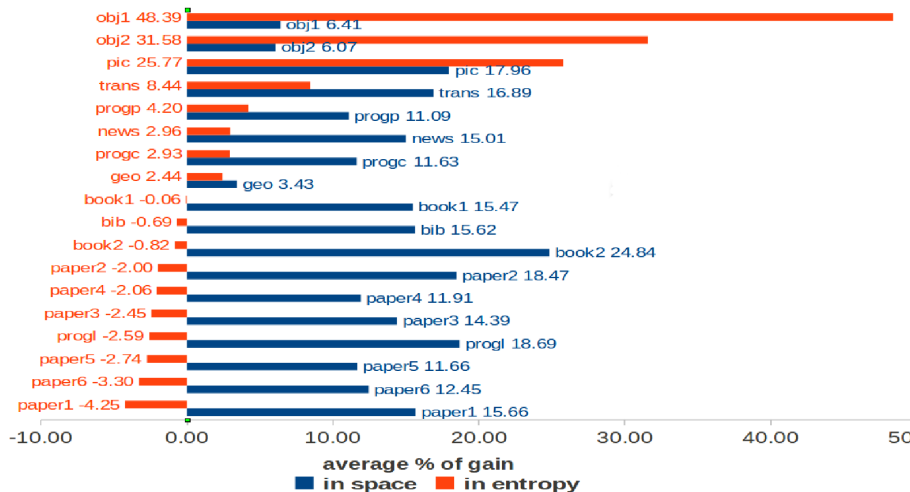
File Name	Percentage of Average Gain in Space	Percentage of Average Gain in Empirical Entropy
paper1	15.66	-4.25
paper6	12.45	-3.30
paper5	11.66	-2.74
progl	18.69	-2.59
paper3	14.39	-2.45
paper4	11.91	-2.06
paper2	18.47	-2.00
book2	24.84	-0.82
bib	15.62	-0.69
book1	15.47	-0.06
geo	3.43	2.44
progc	11.63	2.93
news	15.01	2.96
progp	11.09	4.20
trans	16.89	8.44
pic	17.96	25.77
obj2	6.07	31.58
obj1	6.41	48.39
Average	13.76	5.88

Overall Space–Restricted Comparisons



Overall Space–Restricted Comparisons

5.88% better empirical entropy in 13.76% less space



Conclusions

- A new context modeling technique changing the definition of context in text compression.
- Approximately equal amount of information on encoding of each symbol!
- Enhanced empirical entropy within less space.
- Much better modeling especially on semi-structured or unstructured inputs.

Conclusions

Data compression performance metrics

- Compression ratio (*DePlumP*)
- Compression speed (*PPMII*)
- Resource (memory) requirement

Any previous PPM scheme, as well as new schemes, can be studied with compressed context modeling for possible improvements.

Conclusions

Data compression performance metrics

- Compression ratio (*DePlumP*)
- Compression speed (*PPMII*)
- Resource (memory) requirement

Any previous PPM scheme, as well as new schemes, can be studied with compressed context modeling for possible improvements.

Elegance

Black-art : Too many parameters to be tuned according to the input. Worse, sometimes tuning methods are not well-defined and mostly empirical.

CCM aims to avoid black-art by using a simple metric as information content.

Language modeling

... United ?

¹Sample sentence from CACM Feb'11 issue, DOI:10.1145/1897816.1897842

Language modeling

... United ?
... are amongst the most popular sports in the United ?

¹Sample sentence from CACM Feb'11 issue, DOI:10.1145/1897816.1897842

Language modeling

... United ?

... are amongst the most popular sports in the United ?

Cricket and rugby are amongst the most popular sports in the United ? ¹

¹Sample sentence from CACM Feb'11 issue, DOI:10.1145/1897816.1897842

Language modeling

... United ?

... are amongst the most popular sports in the United ?

Cricket and rugby are amongst the most popular sports in the United ? ¹

- **Power-law (Zipf'1932)** scaling: Small number of words occur frequently and large number of words occur rarely.
- Compression represents frequent items with less bits
- If we consider preceding information content at a particular position, compressed context modeling might be helpful in catching distant relations ???

¹Sample sentence from CACM Feb'11 issue, DOI:10.1145/1897816.1897842

Some applications of language modeling

Sentence generation

Given a word list, which permutation is most plausible ? (e.g., in machine translation)

{toys, with, enjoy, kids, playing} → "Kids enjoy playing with toys."

Author/speaker identification

Develop separate models according to distinct authors/speakers.
Find who generated a given a text/speech.

Space-preserving language modeling with CCM

- Large memory requirement is a serious issue in LM applications.
- Space preserving structure of compressed modeling might be helpful in resource limited environments such the mobile platforms.
 - Speech recognition
 - T9 dictionaries
 - Data compression on mobile data transmission

Variable-Order Markov Models

- Having roots at *J. Rissanen, A universal data compression system, IEEE Trans. on Information Theory, 29-5, 1983*
- Hierarchical HMM (*Fine et al. 1998*), variable length Markov chain (*Bühlman&Wyner, 1999*), probabilistic suffix trees (*Bejenaro&Yona 2001*)
- See Begletier et. al. *On prediction using variable order Markov models, JAIR, vol. 22, 2004.*
- CCM is yet another alternative ?
- If we can achieve better, there will be many applications in diverse fields, e.g., from language/speech applications to bioinformatics (protein classification).

Future Work

- Design and implementation of PPM compressors based on [compressed context modeling](#)
- Exploration of possible improvements on variable–depth Markov modeling via CCM
- Applications on diverse fields using Markov chains

To predict is one thing...

To predict **correctly** is another.



Scientists from the RAND Corporation have created this model to illustrate how a "home computer" could look like in the year 2004. However the needed technology will not be economically feasible for the average home. Also the scientists readily admit that the computer will require not yet invented technology to actually work, but 20 years from now scientific progress is expected to solve these problems. With teletype interface and the Fortran language, the computer will be easy to use.

THANK YOU!