Контекстное моделирование

- Й. Риссаннен
- Г. Лэнгдон
- 1981г.

- Процесс сжатия сообщения реализуется в два этапа
  - моделирование источника
  - кодирование (в соответствии с полученной моделью)

- Практически всегда истинная статистика источника скрыта,
- поэтому для кодирования очередного символа сообщения необходимо строить вероятностное распределение или модель источника

- Чем точнее построенная модель (или предсказание очередного символа),

  тем меньше избыточность построенного на основе этой модели кода.

- Техника контекстного моделирования Prediction by Partial Matching (PPM)

Клири, Уиттен 1984 г.

- Контекстное моделирование –

оценивание вероятности появления текущего символа в зависимости от контекста.

Контекст – последовательность предыдущих символов сообщения

- Контекстная модель порядка m

  КМ(m) – это набор оценок вероятностей символов, которые строятся на основании обычных счетчиков частот, связанных с текущим

  контекстом длины m.

КЛОКОКЛ

	К	О	Л
КМ(-1)	1	1	1
КМ(0)	3	2	2
КМ(1) контекст «К»	-	1	2
КМ(2) контекст «ОК»	-	1	1

- Оценка вероятности появления символа X в контексте S

$$P(X \mid S) = \frac{N(X \mid S)}{N(S)}$$

где N(X|S) – количество появлений символа X после контекста S

N(S) – общее количество подстрок S

КЛОКОКЛ

	К	О	Л
КМ(-1)	1/3	1/3	1/3
КМ(0)	3/7	2/7	2/7
КМ(1) контекст «К»	-	1/3	2/3
КМ(2) контекст «ОК»	-	1/2	1/2

- В случае если контекст еще не встречался в сообщении или очередной символ не появлялся в данном контексте, то нельзя оценить вероятность
- Для устранения такой ситуации в алфавит вводят дополнительный специальный символ.

- Символ ухода – виртуальный символ, который показывает необходимость перехода к контекстной модели более низкого порядка.
- Для кодирования символа ухода также необходимо оценить вероятность его использования.

- Вероятность ухода – это суммарная вероятность всех символов алфавита входного потока, еще ни разу не появлявшихся в контексте.
- Любая КМ должна давать отличную от нуля оценку вероятности ухода. Кроме случаев, когда все оценки вероятностей символов отличны от нуля.

- Оценка вероятности появления символа X в контексте S с учетом символа ухода

$$P(X \mid S) = \frac{N(X \mid S)}{N(S) + 1}$$

где N(X|S) – количество появлений символа X после контекста S

N(S) – общее количество подстрок S

Схема метода PPM

- 1. Если вероятность очередного символа X оценивается КМ(N) ненулевым числом, то код строится с использованием этой оценки.
- 2. Иначе выдается сигнал в виде символа ухода и происходит переход к шагу 1 с КМ(N-1) до тех пор пока символ X не будет оценен ненулевым числом.

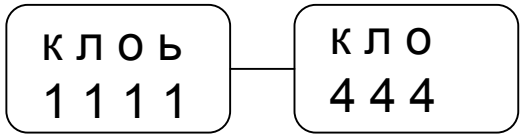
- КМ(-1) гарантирует, что вероятность очередного символа X будет оценена.
- Если в процессе оценки обнаруживается, что текущий рассматриваемый контекст встречается первый раз, то для него создается КМ.
- После кодирования символа X происходит обновление статистики всех КМ, которые использовались при оценке.

клококлолкол?

к	л	о	ь
1	1	1	1

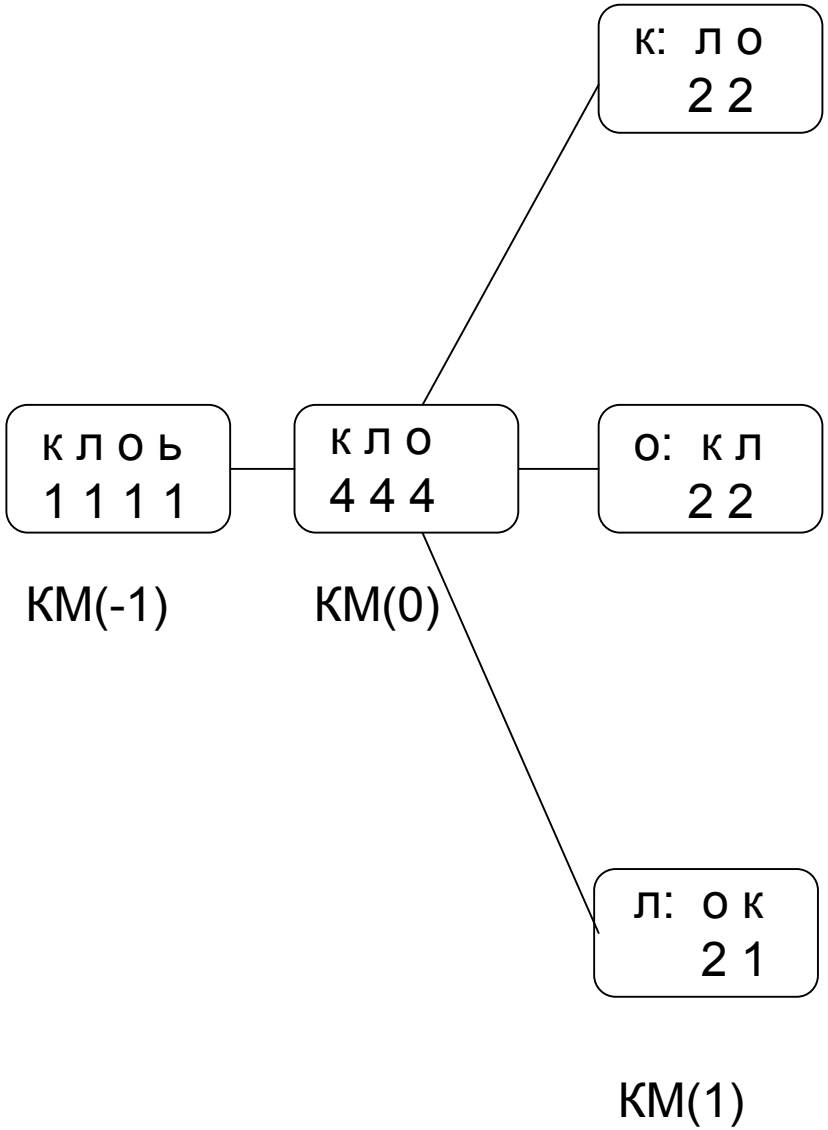
КМ(-1)

клококлолкол?



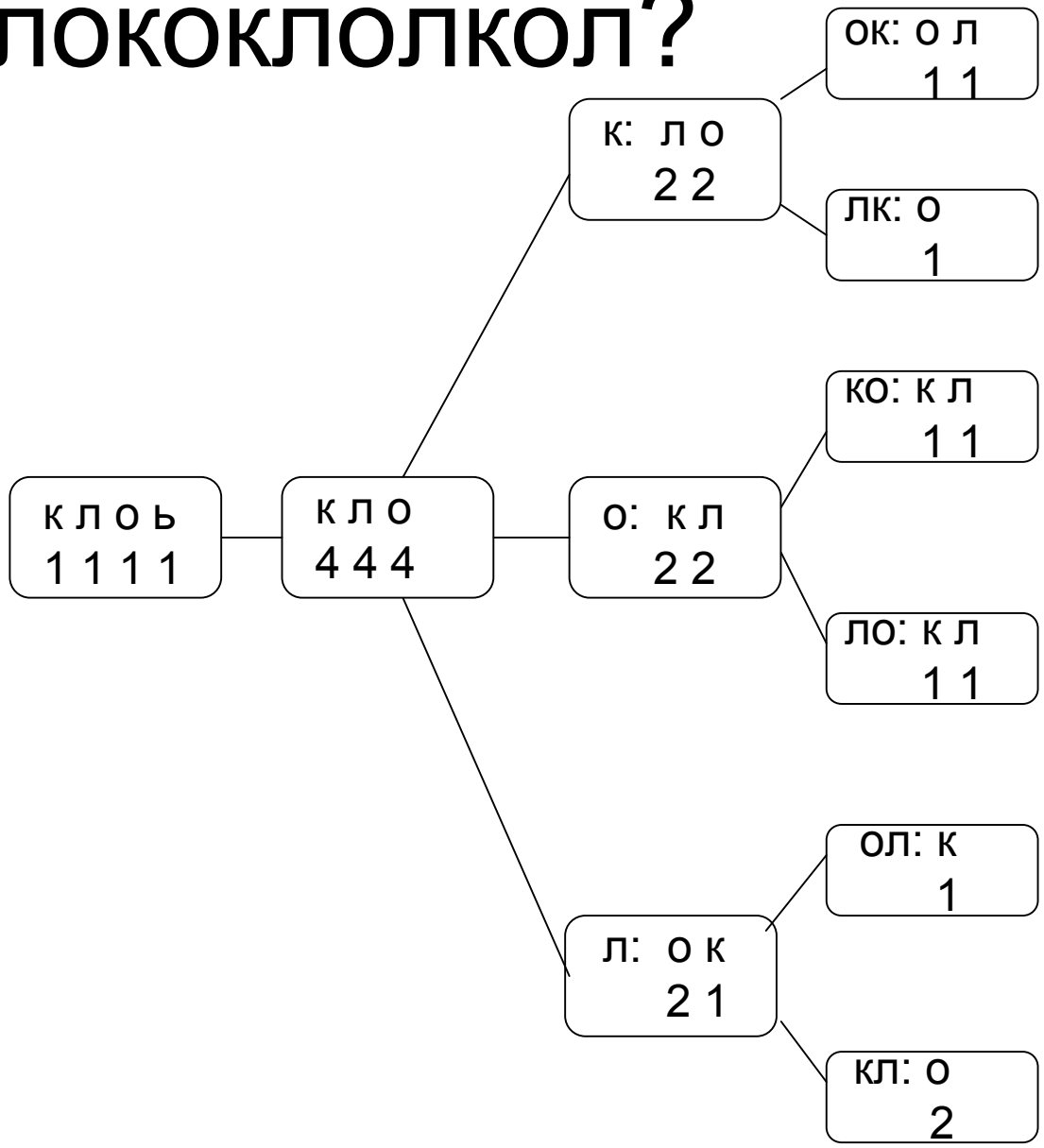
КМ(-1) КМ(0)

клококлолкол?



КМ(2)

клококлолкол?



КМ(3)

клококлолкол?

Оценки вероятностей

	3	2	1	0	-1	оценка
	«кол»	«ол»	«л»	«»		
к	-	1/(1+1)	-	-	-	1/2
л	-	1/(1+1)	1/(2+1)	4/(4+1)	-	2/15
о	-	1/(1+1)	2/(2+1)	-	-	1/3
ь	-	1/(1+1)	1/(2+1)	1/(4+1)	1	1/30

КМ(3)

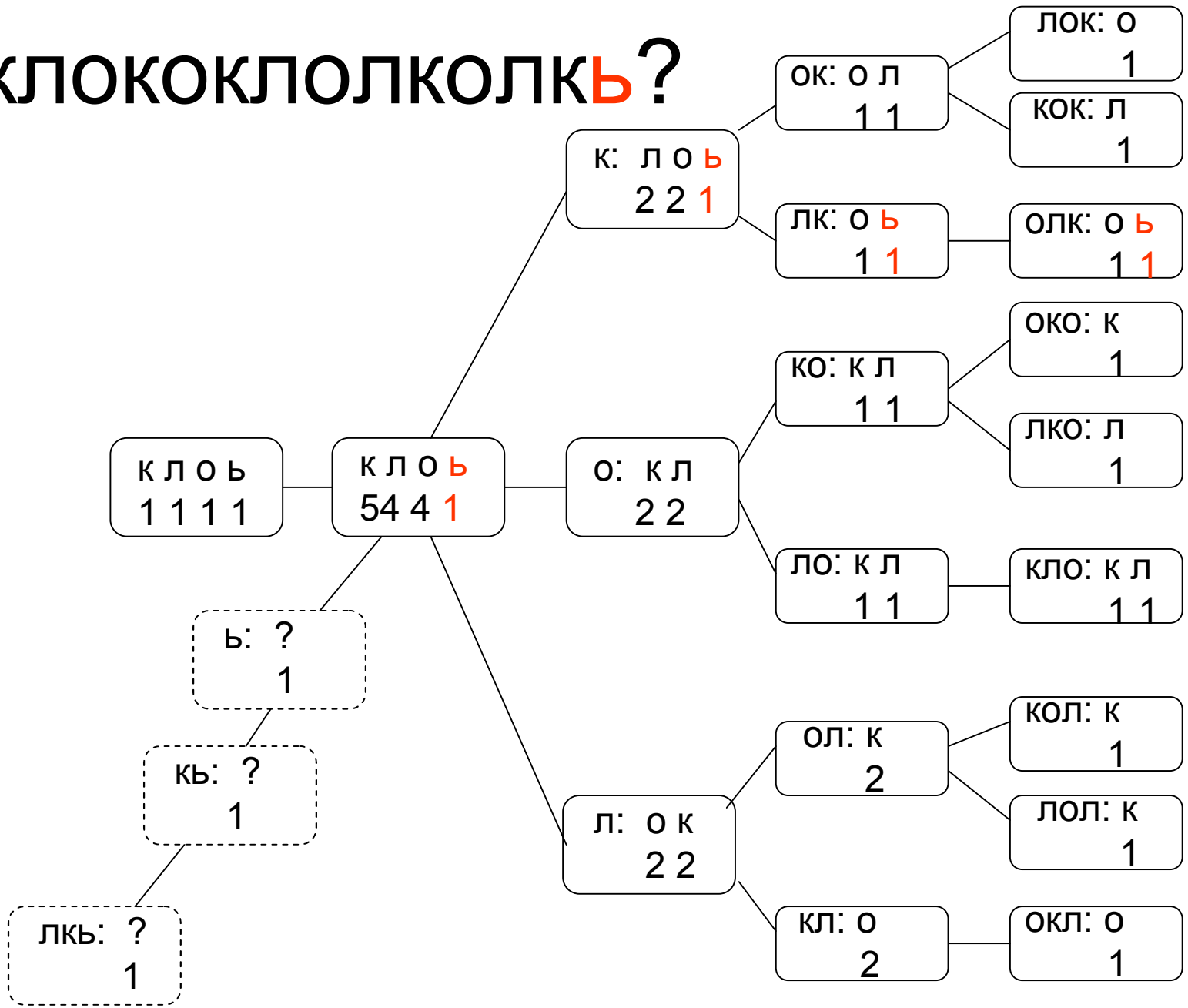
клококлолколк?

Оценки вероятностей

	3	2	1	0	-1	оценка
	«олк»	«лк»	«к»	«»		
к	1/(1+1)	1	1/(2+1)	5/(5+1)	-	5/36
л	1/(1+1)	1	2/(2+1)	-	-	1/3
о	1/(1+1)	-	-	-	-	1/2
ь	1/(1+1)	1	1/(2+1)	1/(5+1)	1	1/36

КМ(3)

клококлолколкь?



Оценки вероятностей

	3	2	1	0	-1	оценка
	«лкь»	«кь»	«ь»	«»		
к	-	-	-	5/14	-	5/14
л	-	-	-	4/14	-	4/14
о	-	-	-	4/14	-	4/14
ь	-	-	-	1/14	-	1/14