DL: Машинный перевод

План

- Статистический машинный перевод
- Эмбединги
- Нейронные сети

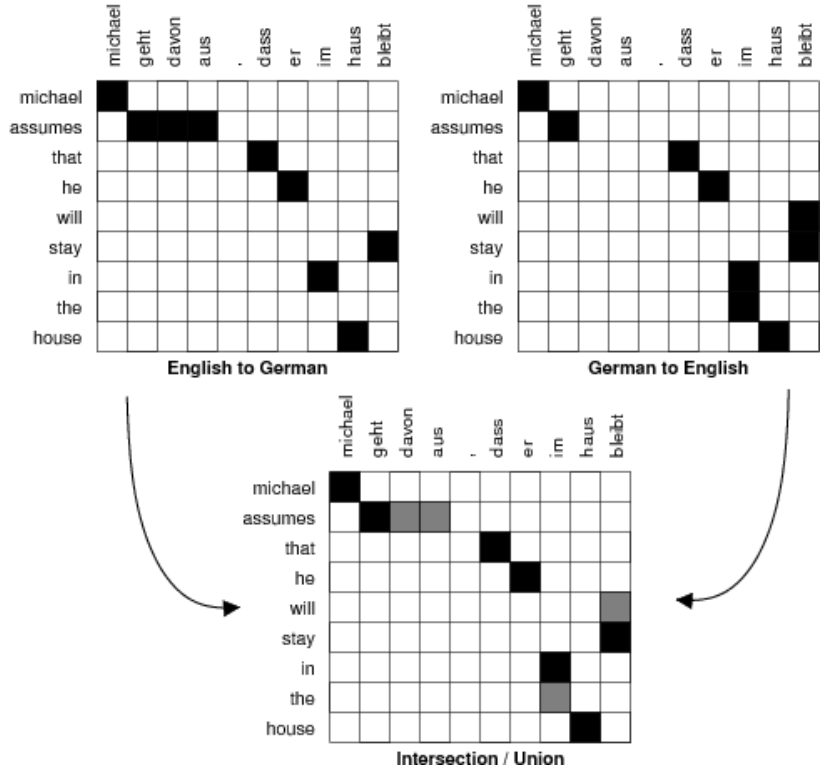
Статистический машинный перевод

$$argmax_y P(y|x)$$

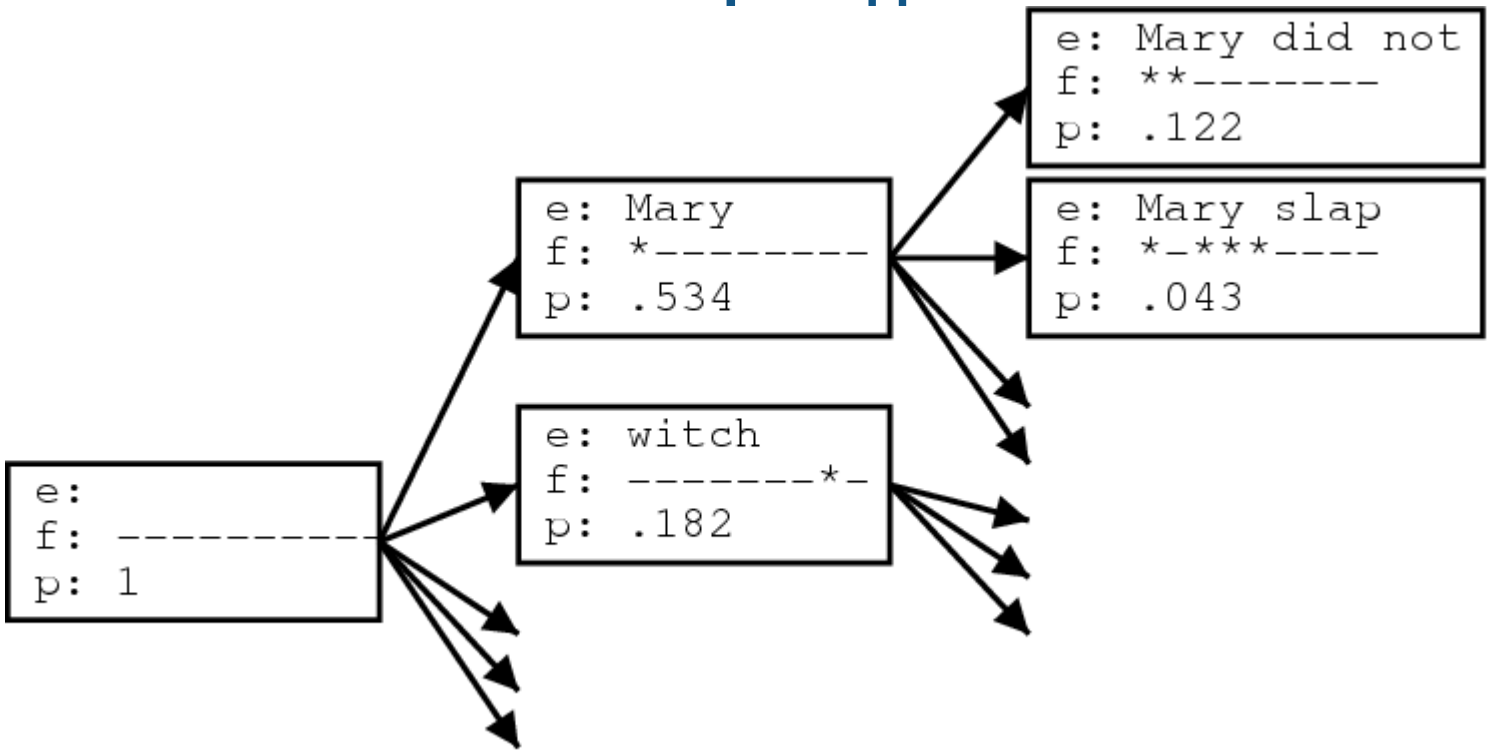
$$argmax_y P(x|y)P(y)$$

$$P(x,a|y)P(y)$$

Статистический машинный перевод: alignment

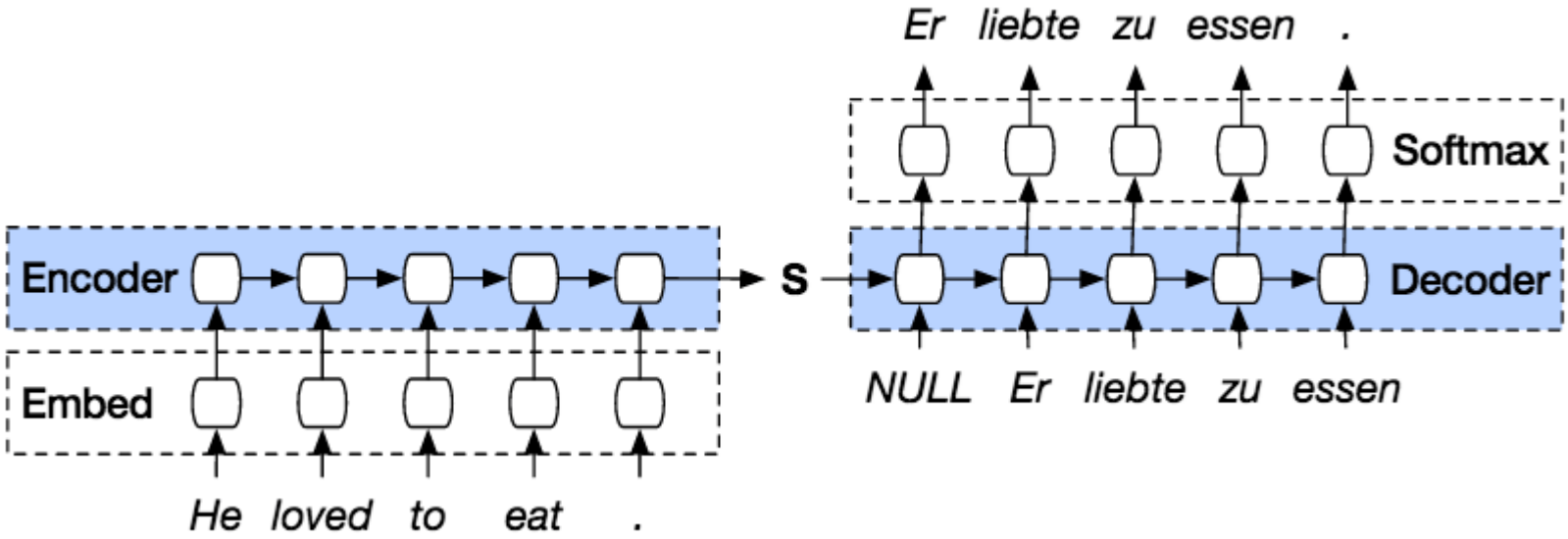


Статистический машинный перевод

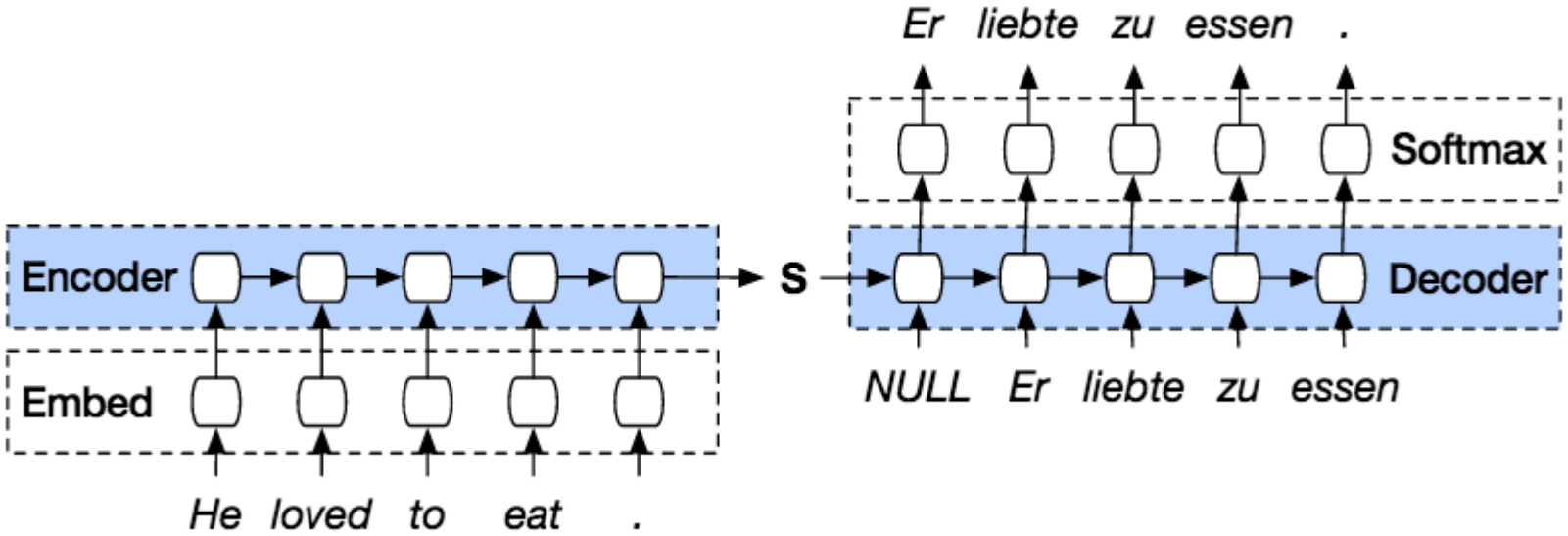


Статистический машинный перевод

Seq2seq 2014

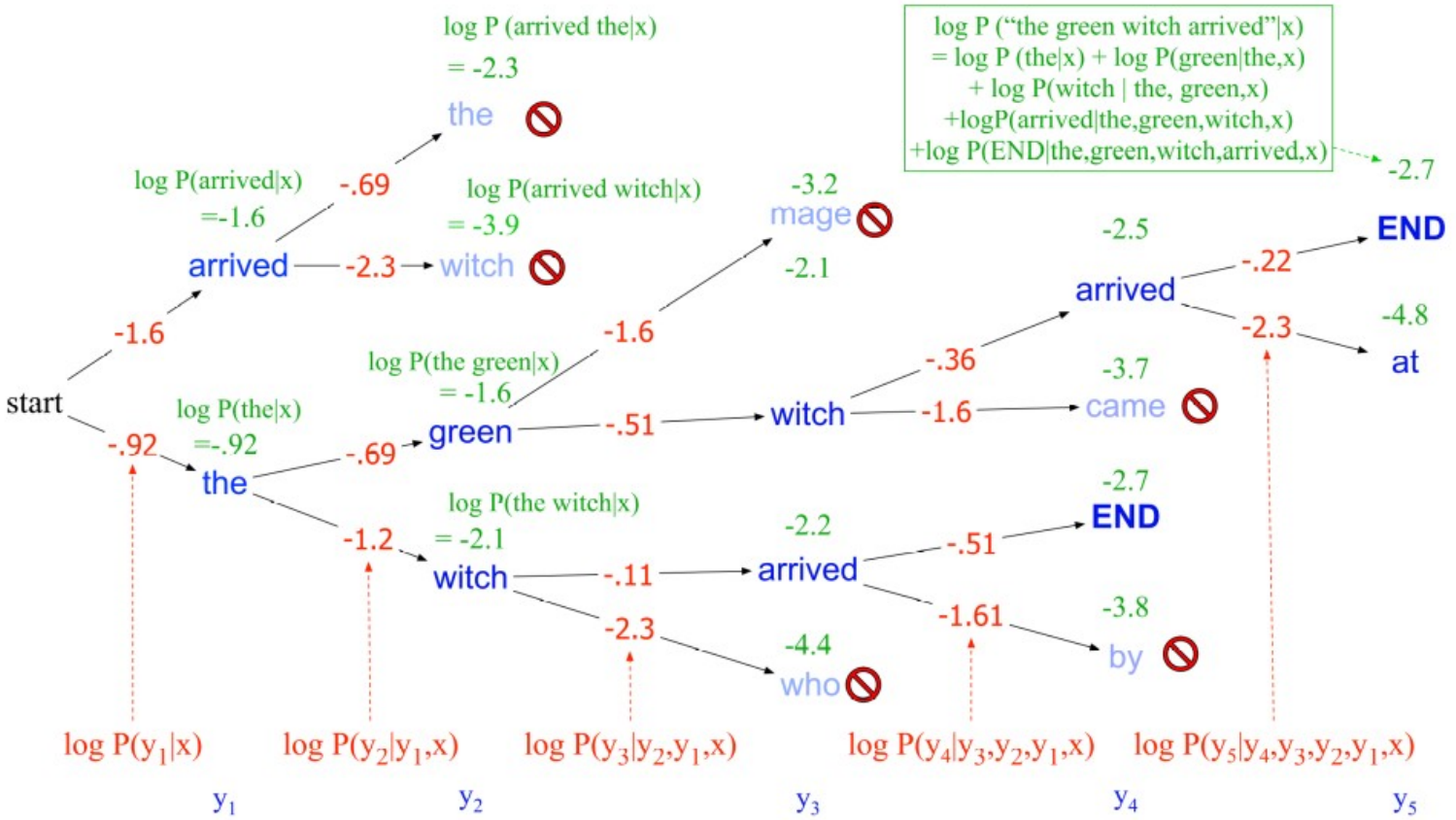


Seq2seq greedy vs beam search



$$score(y_1, y_2, \ldots, y_n) = \log P(y_1, y_2, \ldots, y_n | x) = \sum_{i=1}^{n} \log P(y_i | y_1, \ldots, y_{i-1} | x)$$

beam search


log P (arrived the|x)
= -2.3
the
log P(arrived|x)
=-1.6
-.69
log P(arrived witch|x)
= -3.9
arrived
-2.3
witch
-1.6
log P(the green|x)
= -1.6
-3.2
mage
-2.1
start
log P(the|x)
=-.92
-.92
the
-.69
green
-.51
witch
-.36
arrived
-2.5
-.22
END
-2.3
-4.8
at
-1.6
-3.7
came
log P(the witch|x)
= -2.1
-1.2
witch
-.11
-2.2
arrived
-.51
-2.7
END
-2.3
-1.61
-3.8
by
-4.4
who
log P ("the green witch arrived"|x)
= log P (the|x) + log P(green|the,x)
+ log P(witch | the, green,x)
+logP(arrived|the,green,witch,x)
+log P(END|the,green,witch,arrived,x)
-2.7
log P(y1|x)
log P(y2|y1,x)
log P(y3|y2,y1,x)
log P(y4|y3,y2,y1,x)
log P(y5|y4,y3,y2,y1,x)
y1
y2
y3
y4
y5

beam search: нормировка

$$score(y_1, y_2, \dots, y_n) = \log P(y_1, y_2, \dots, y_n | x) = \sum_{i=1}^{n} \log P(y_i | y_1, \dots, y_{i-1} | x)$$

$$\frac{1}{n} \sum_{i=1}^{n} \log P(y_i | y_1, \dots, y_{i-1} | x)$$

BLEU

- N-gram overlap between machine translation output and reference translation
- Compute precision for n-grams of size 1 to 4
- Add brevity penalty (for too short translations)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

- Typically computed over the entire corpus, not single sentences

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

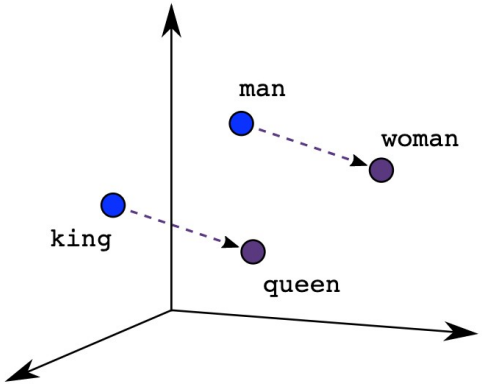
SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

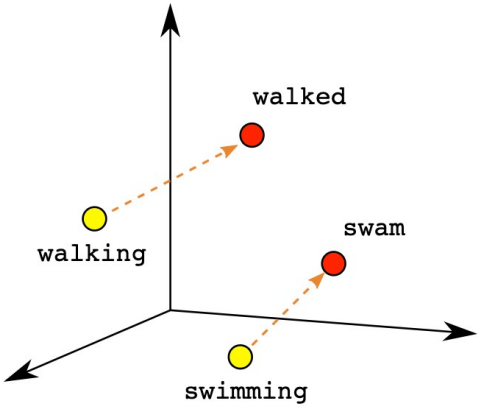
Word Embedding: one hot

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

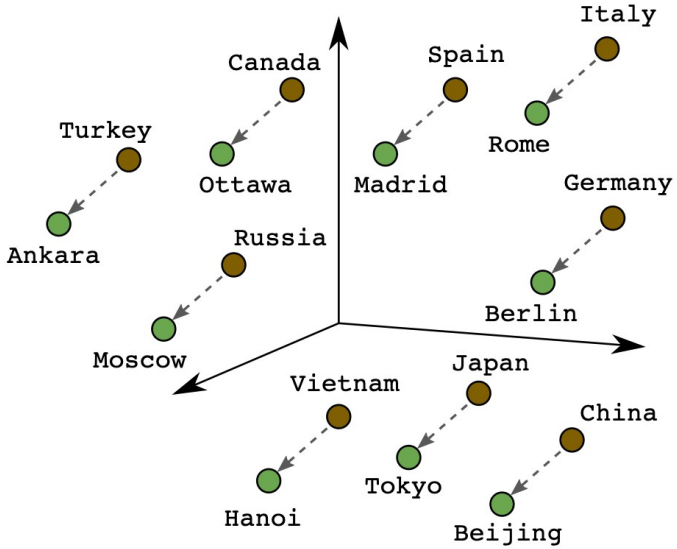
Word Embedding: word2vec



Male-Female



Verb Tense



Country-Capital

Word Embedding: word2vec