Neural Nets in Machine Translation

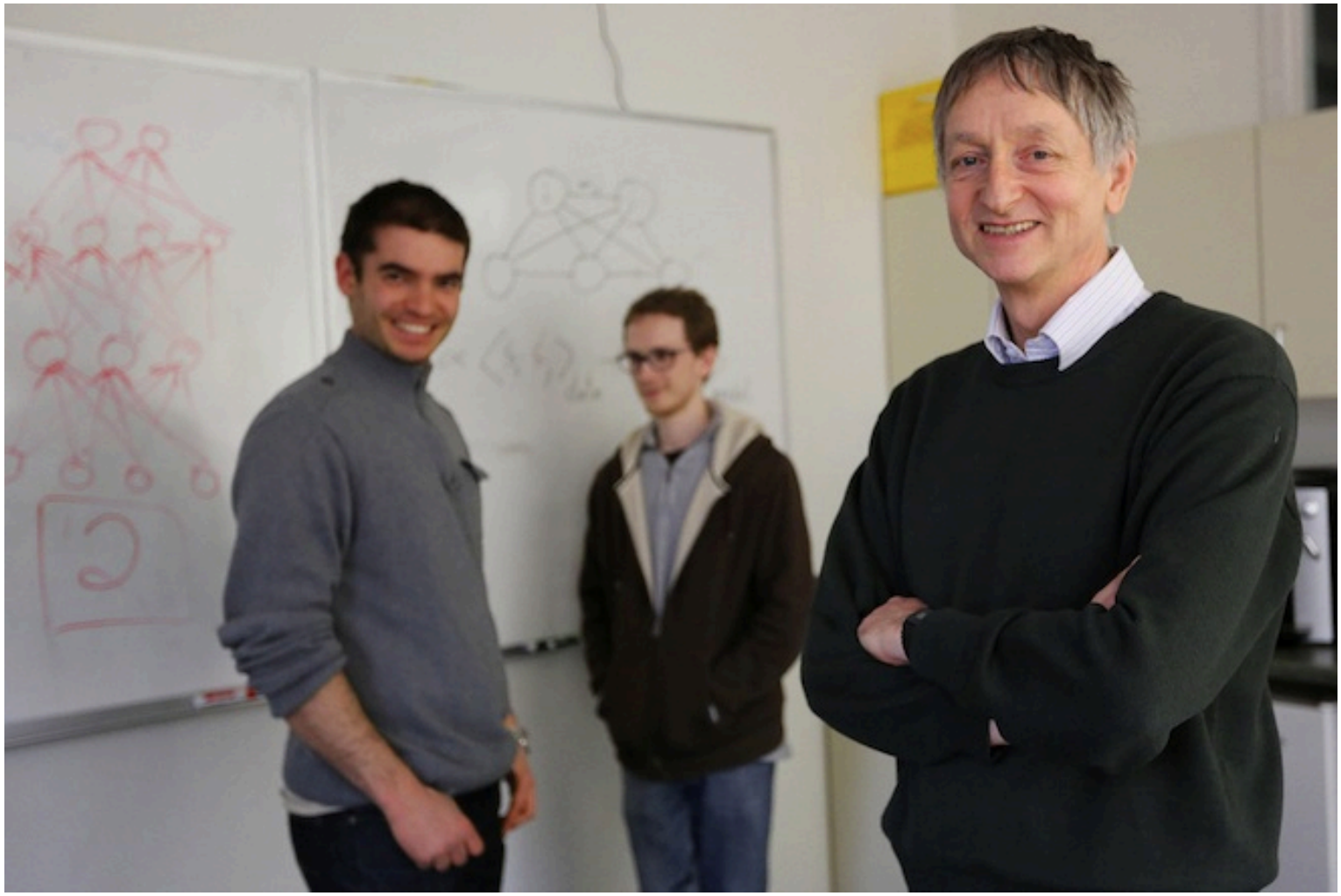
Alexander Fonarev
newo@newo.su

March 24, 2014

History of neural nets and deep learning

- Invented – 1960's
- Backpropagation – 1974
- RBMs and DBMs – 2000
- Autoencoders – 2006
- Dropout – 2012

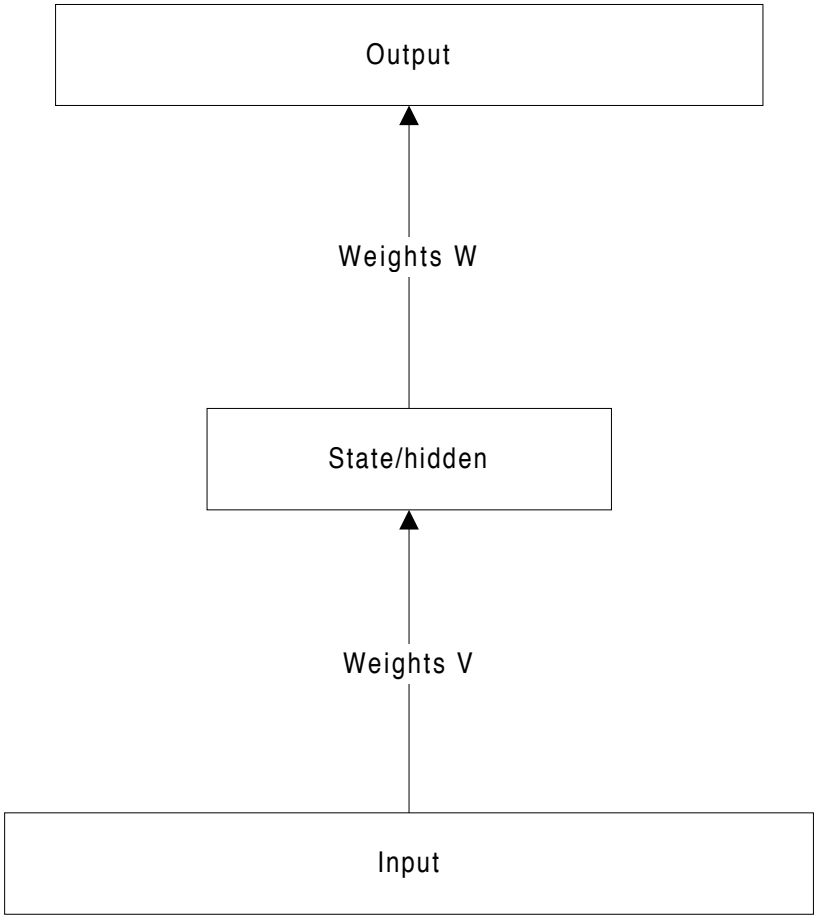
Geoffrey Hinton



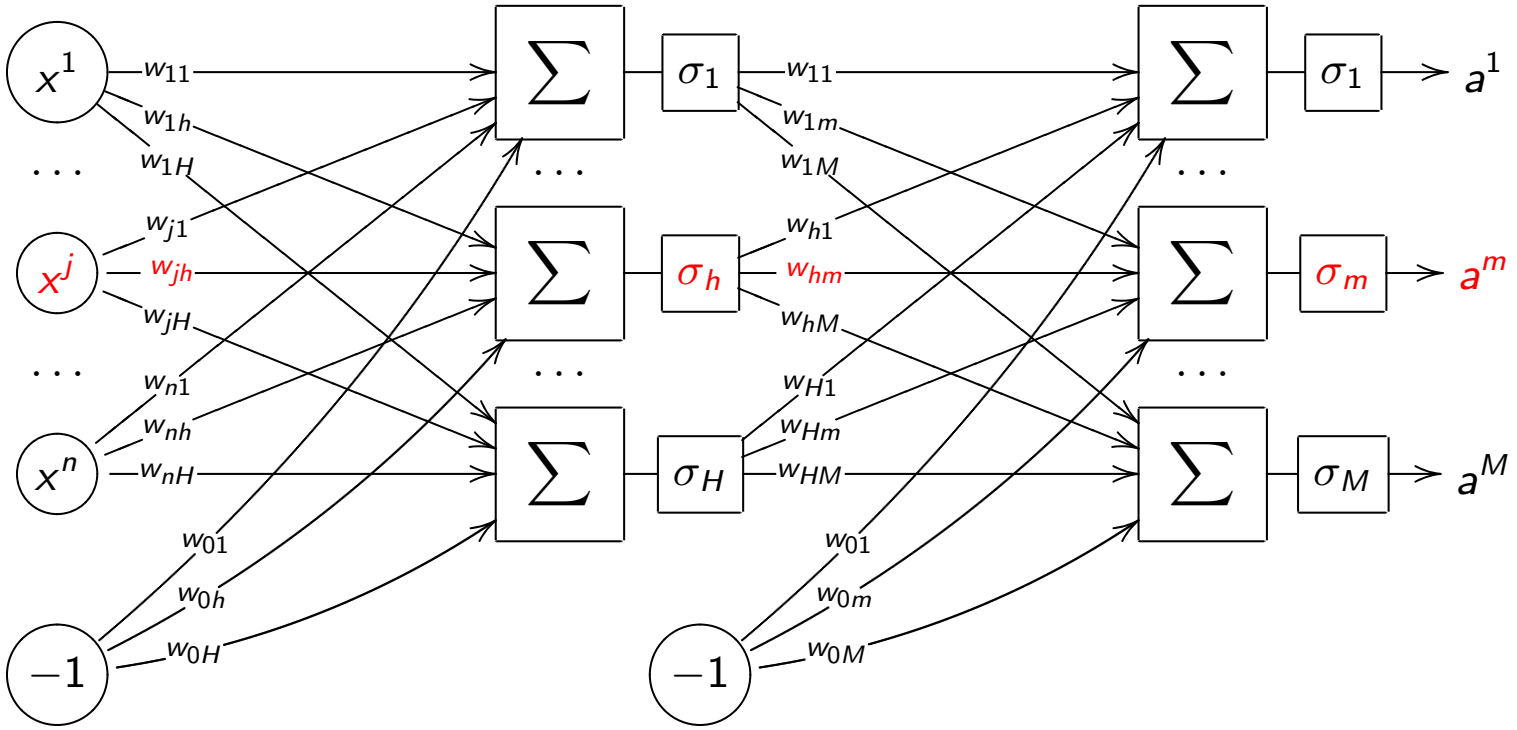
Whetting your appetite

- Win different competitions since 2009
- Hinton's team is bought by Google in 2013
- State of the art now
  - Image recognition
  - Speech recognition
  - Sentiment analysis
  - Paraphrase detection
  - etc

Feed forward neural net



Feed forward neural net

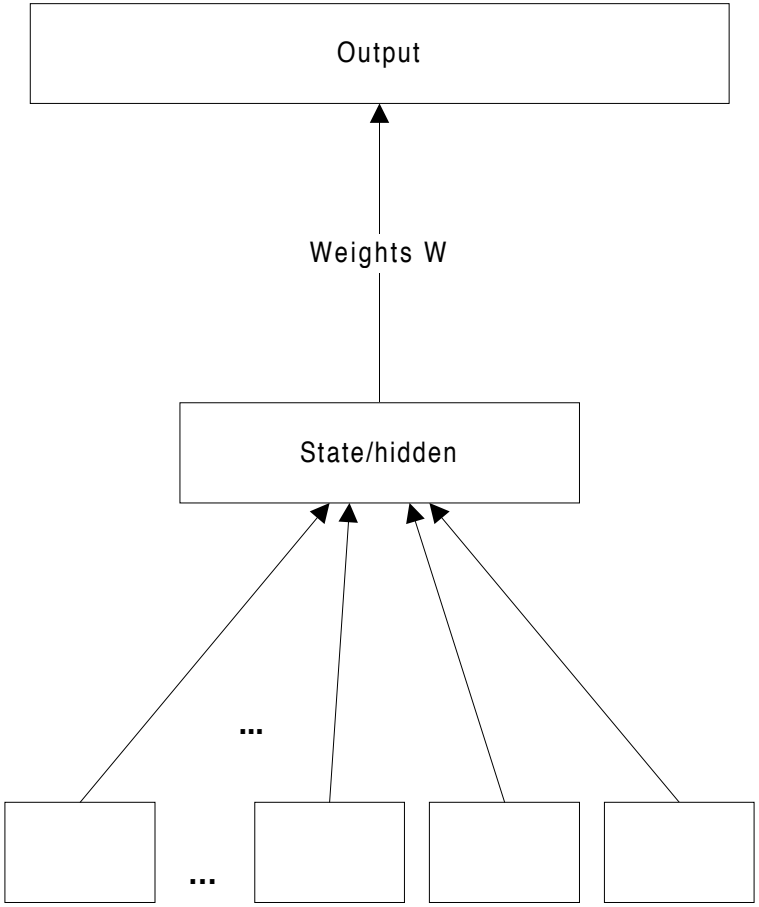


Language modeling

$$\hat{P}(w_1^T) = \prod_{t=1}^{T} \hat{P}(w_t | w_1^{t-1})$$

$$\hat{P}(w_t | w_1^{t-1}) \approx \hat{P}(w_t | w_{t-n+1}^{t-1})$$

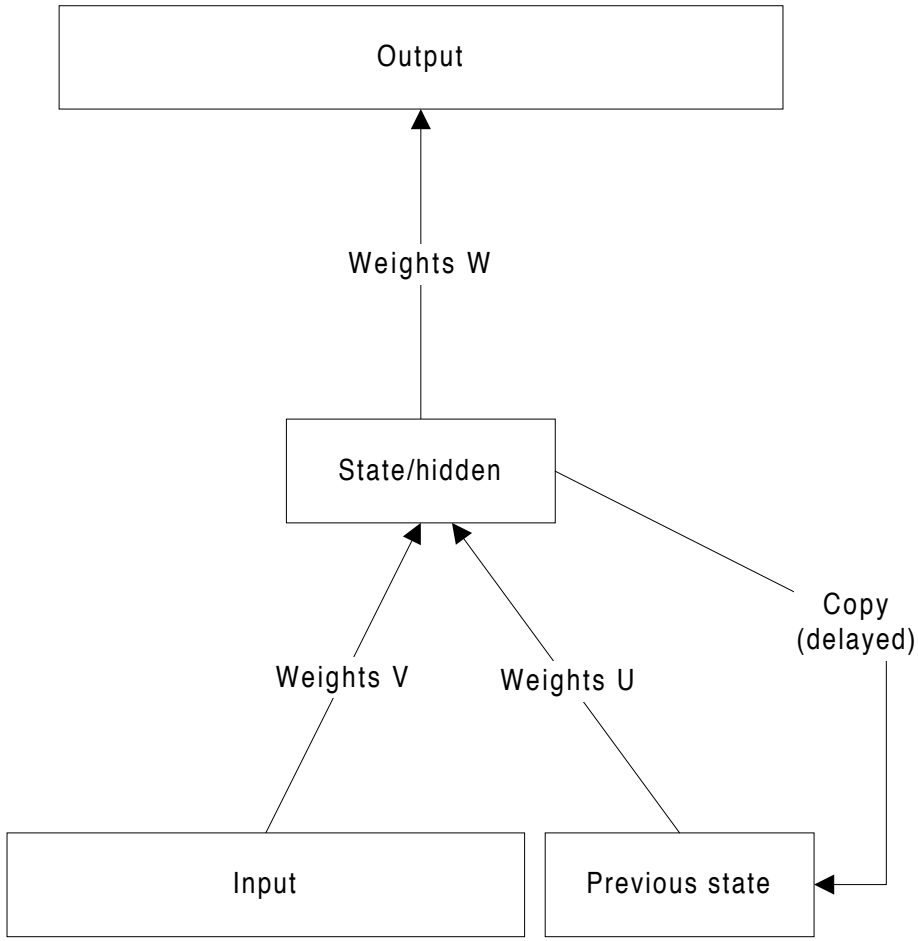
Neural network language model



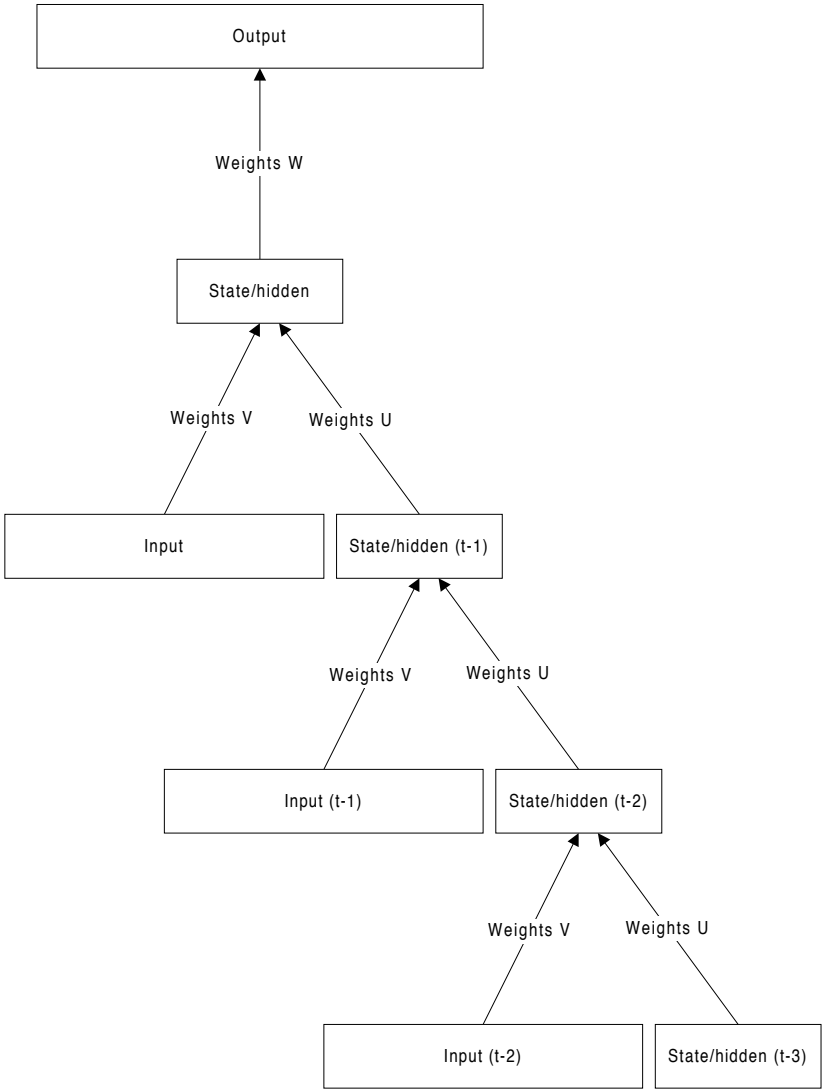
Recurrent neural net

$$\hat{P}(w_1^T) = \prod_{t=1}^{T} \hat{P}(w_t | w_1^{t-1})$$

$$\hat{P}(w_t | w_1^{t-1}) \approx \hat{P}(w_t | w_{t-n+1}^{t-1})$$

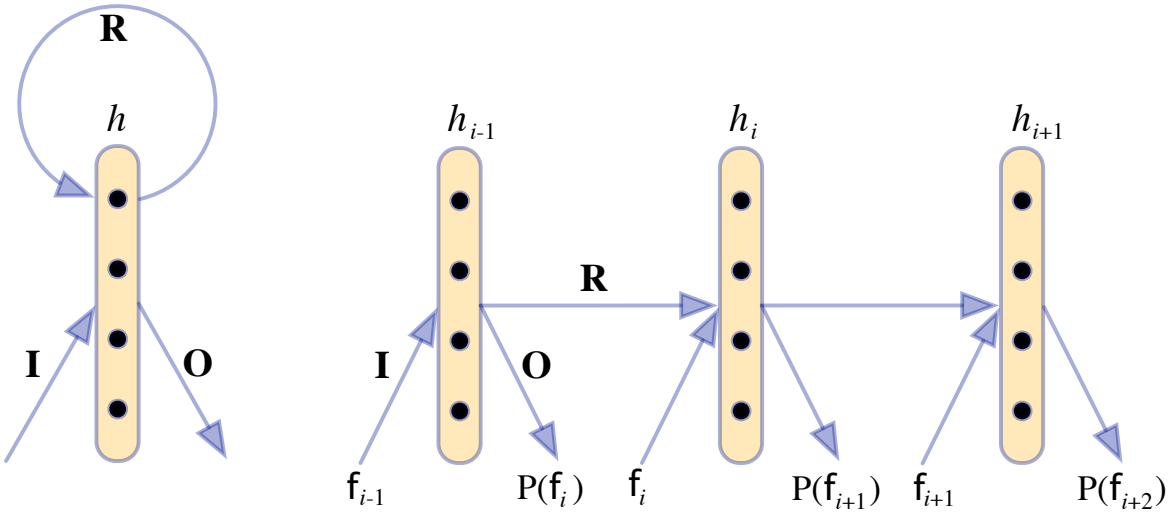


Unfolded recurrent neural net



Recurrent neural network LM

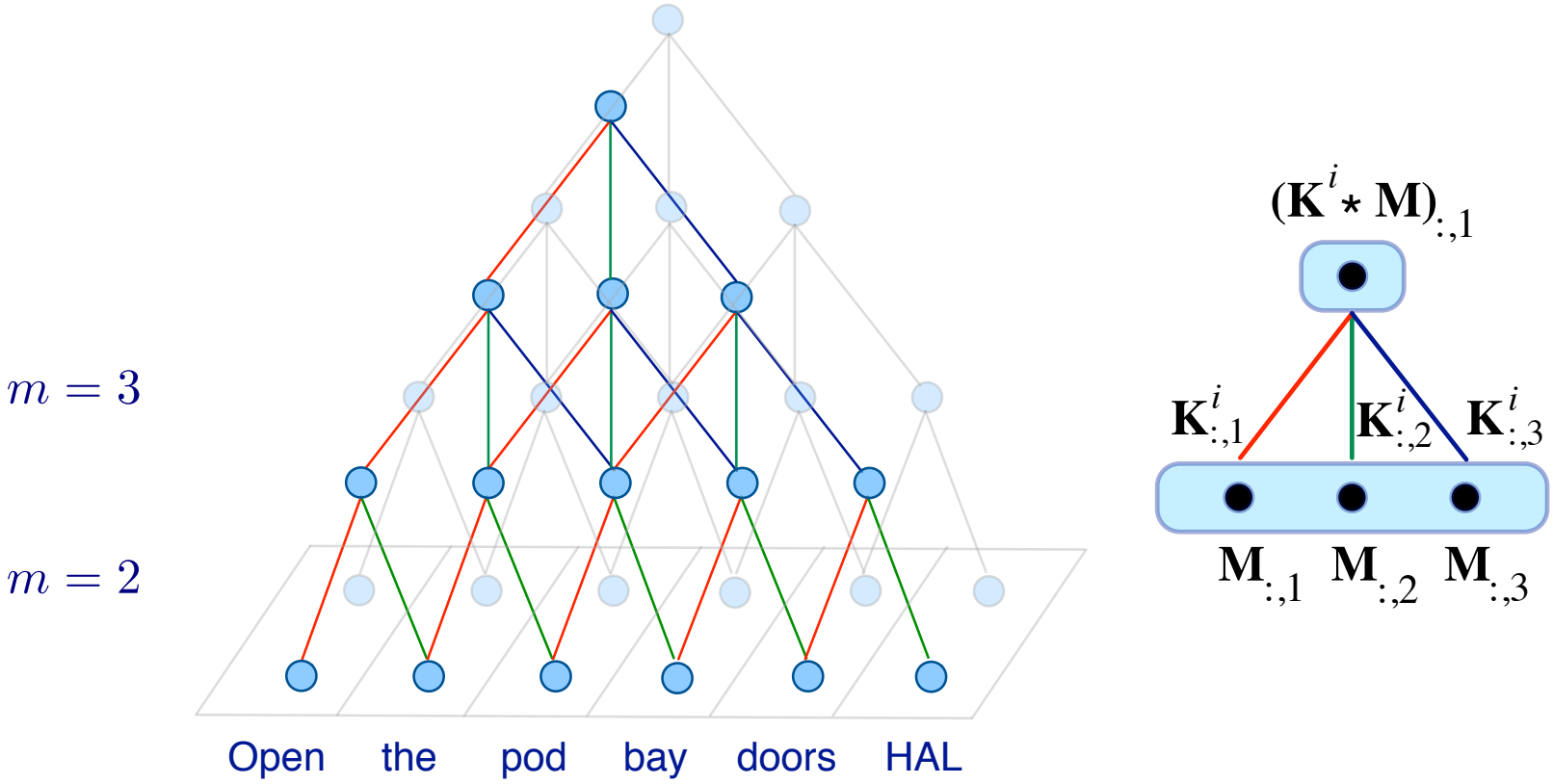
$$h_1 = \sigma(\mathbf{I} \cdot \mathsf{v}(\mathsf{f}_1))$$
$$h_{i+1} = \sigma(\mathbf{R} \cdot h_i + \mathbf{I} \cdot \mathsf{v}(\mathsf{f}_{i+1}))$$
$$o_{i+1} = \mathbf{O} \cdot h_i$$



Recurrent neural network LM

Model	WER static	WER dynamic
RT05 LM	24.5	-
RT09 LM - baseline	24.1	-
KN5 in-domain	25.7	-
RNN 500/10 in-domain	24.2	24.1
RNN 500/10 + RT09 LM	**23.3**	23.2
RNN 800/10 in-domain	24.3	23.8
RNN 800/10 + RT09 LM	23.4	23.1
RNN 1000/5 in-domain	24.2	23.7
RNN 1000/5 + RT09 LM	23.4	22.9
3xRNN + RT09 LM	**23.3**	**22.8**

Convolutional sentence model



$$\mathbf{K}^i_{:,1} \odot \mathbf{M}_{:,a} + \mathbf{K}^i_{:,2} \odot \mathbf{M}_{:,a+1} + \mathbf{K}^i_{:,3} \odot \mathbf{M}_{:,a+2}$$

Recurrent continuous translation model 1

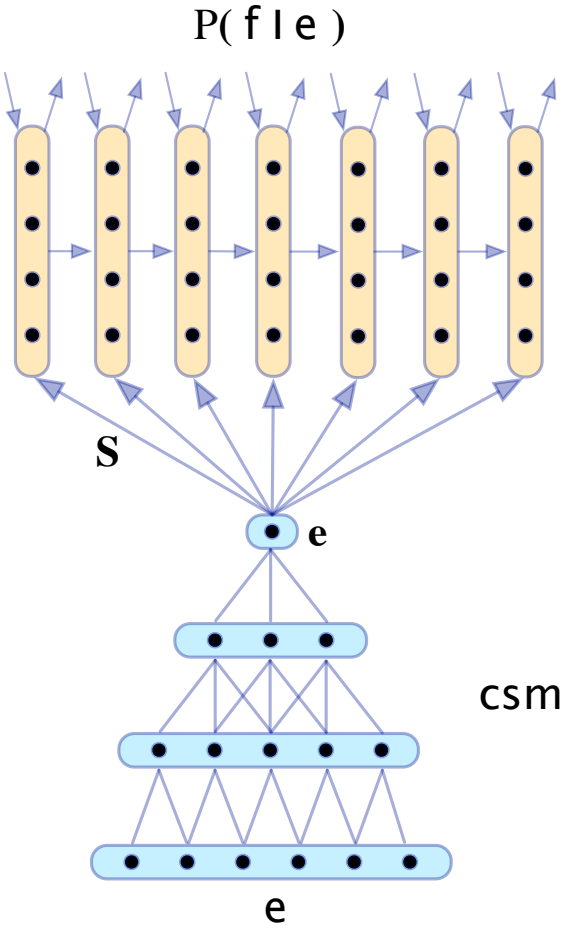
$$P(\mathsf{f}|\mathsf{e}) = \prod_{i=1}^{m} P(\mathsf{f}_i|\mathsf{f}_{1:i-1}, \mathsf{e})$$

$$\mathbf{s} = \mathbf{S} \cdot \mathsf{csm}(\mathsf{e})$$

$$h_1 = \sigma(\mathbf{I} \cdot \mathsf{v}(\mathsf{f}_1) + \mathbf{s})$$

$$h_{i+1} = \sigma(\mathbf{R} \cdot h_i + \mathbf{I} \cdot \mathsf{v}(\mathsf{f}_{i+1}) + \mathbf{s})$$

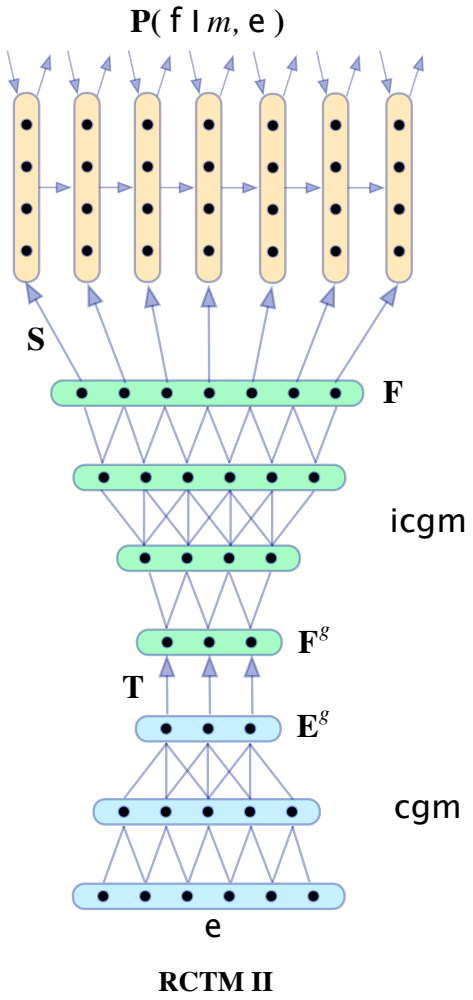
$$o_{i+1} = \mathbf{O} \cdot h_i$$



Recurrent continuous translation model 2

$$P(\mathsf{f}|\mathsf{e}) = P(\mathsf{f}|m,\mathsf{e}) \cdot P(m|\mathsf{e})$$
$$= \prod_{i=1}^{m} P(\mathsf{f}_{i+1}|\mathsf{f}_{1:i}, m, \mathsf{e}) \cdot P(m|\mathsf{e})$$

$$\mathbf{E}^g = \mathsf{cgm}(\mathsf{e}, 4)$$
$$\mathbf{F}^g_{:,j} = \sigma(\mathbf{T} \cdot \mathbf{E}^g_{:,j})$$
$$\mathbf{F} = \mathsf{icgm}(\mathbf{F}^g, m)$$
$$h_1 = \sigma(\mathbf{I} \cdot \mathsf{v}(\mathsf{f}_1) + \mathbf{S} \cdot \mathbf{F}_{:,1})$$
$$h_{i+1} = \sigma(\mathbf{R} \cdot h_i + \mathbf{I} \cdot \mathsf{v}(\mathsf{f}_{i+1}) + \mathbf{S} \cdot \mathbf{F}_{:,i+1})$$
$$o_{i+1} = \mathbf{O} \cdot h_i$$



Perplexity

WMT-NT	*2009*	*2010*	*2011*	*2012*
KN-5	218	213	222	225
RLM	178	169	178	181
IBM 1	207	200	188	197
FA-IBM 2	153	146	135	144
RCTM I	143	134	140	142
RCTM II	**86**	**77**	**76**	**77**

Perplexity on permuted data

WMT-NT PERM	*2009*	*2010*	*2011*	*2012*
RCTM II	174	168	175	178

How to translate?

Generate words from distribution ☹

WMT-NT	*2009*	*2010*	*2011*	*2012*
RCTM I + WP	19.7	21.1	22.5	21.5
RCTM II + WP	19.8	21.1	22.5	21.7
`cdec` (12 features)	19.9	21.2	22.6	21.8

Examples

English source sentence	French gold translation	RCTM II candidate translation	Rank
the patient is sick .	le patient est malade .	le patient est insuffisante . le patient est mort . la patient est insuffisante .	1 4 23
the patient is dead .	le patient est mort .	le patient est mort . le patient est dépassé .	1 4
the patient is ill .	le patient est malade .	le patient est mal .	3
the patients are sick .	les patients sont malades .	les patients sont confrontés . les patients sont corrompus .	2 5
the patients are dead .	les patients sont morts .	les patients sont morts .	1
the patients are ill .	les patients sont malades .	les patients sont confrontés .	5
the patient was ill .	le patient était malade .	le patient était mal .	2
the patients are not dead .	les patients ne sont pas morts .	les patients ne sont pas morts .	1
the patients are not sick .	les patients ne sont pas malades .	les patients ne sont pas $\langle unknown \rangle$. les patients ne sont pas mal .	1 6
the patients were saved .	les patients ont été sauvés .	les patients ont été sauvées .	6

References

- *A guide to recurrent neural networks and backpropagation, 2001 [Mikael Boden]*
- *A Neural Probabilistic Language Model, 2003 [Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin]*
- *Recurrent neural network based language model, 2010 [Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan "Honza" Cernock, Sanjeev Khudanpur]*
- *Recurrent Continuous Translation Models, 2013 [Nal Kalchbrenner, Phil Blunsom]*
- *Compositional Semantics, Deep Learning, and Machine Translation, MTM 2013 [Phil Blunsom]*
- *Машинное обучение (курс лекций), К.В.Воронцов*