RWTH AACHEN UNIVERSITY

RAAI School 2019:
Summer School of Russian Association for Artificial Intellgence
MIPT campus at Dolgoprudny, Moscow, Russia, July 4-7, 2019

Speech Recognition and Machine Translation:
From Bayes Decision Rule to Deep Learning

Hermann Ney

Lehrstuhl Informatik 6
Human Language Technology and Pattern Recognition
RWTH Aachen University, Aachen, Germany

My Presentation

some of today's buzz words in deep learning for ASR:

- **ANN structures: MLP and (LSTM) RNN**
- **sequence-to-sequence modelling/systems/training**
- **CTC: connectionist temporal classification**
- **end-to-end modelling/systems/training**
- **discriminative modelling/training**
- **...**
- **many of these methods: very successful**

problems:

- **lots of 'noisy' experimental results,**
  of 'unorthodox' ideas and of re-invented/re-named concepts
- **important question:**
  how to filter out the noise in the experimental results?
  what are really fundamental (mathematical) principles that we can rely on?

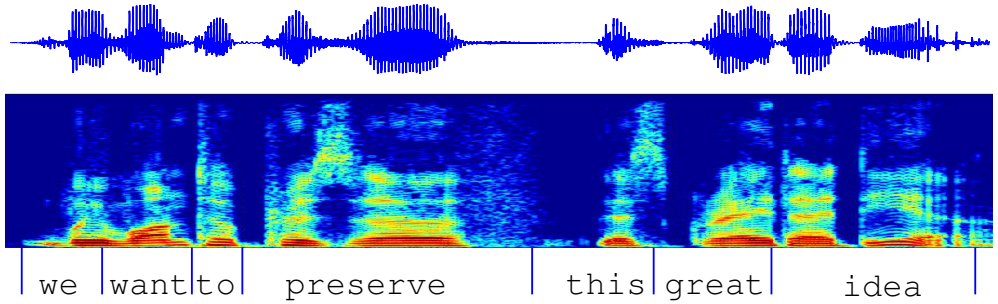
My Presentation

- **my experience over the last 40 years:**
  share my interpretations and views (maybe controversial!)
- **right starting point:**
  principles are given by Bayes decision theory
- **many details that have to be filled in:**
  type of models, training criteria, etc.
- **question:**
  how do deep learning methods fit into Bayes decision theory?
- **tasks in machine learning/statistical classification:**
  – speech recognition
  – machine translation (purely symbolic!) and other tasks in NLP
- **yes, deep learning has been very successful,**
  but there has been, is and will be life outside deep learning!

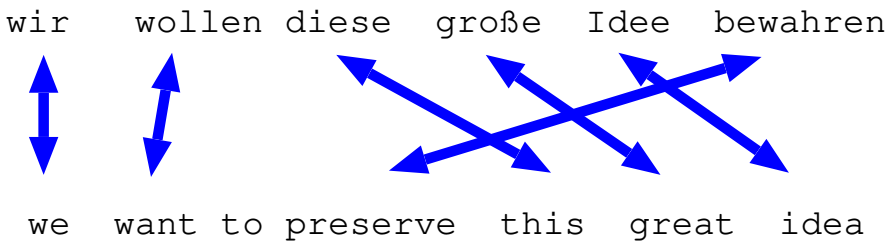
  note: ANN $\stackrel{?}{=}$ matrix-vector product + nonlinearity + concatenation

Tasks in Human Language Technology
Sequence-to-Sequence Processing

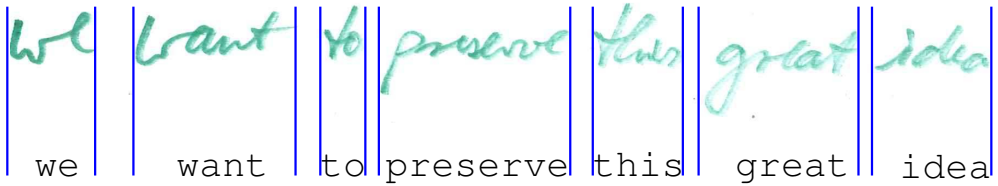
Automatic Speech Recognition (ASR)
(speech signal processing)



Machine Translation (MT)
(symbol or text processing)



Handwriting Recognition (HWR)
(image signal processing)



more tasks:
– sign language (gesture) recognition
– syntactic or semantic tagging (NLU)
– ...

Projects on ASR and MT

RWTH AACHEN UNIVERSITY

projects (many large-scale joint projects):

- **SPICOS 1984-89**
- **Verbmobil 1993-2000**
- **TC-STAR 2004-2007: funded by EU**
  - **partners: FBK, LIMSI, KIT, UPC, RWTH, IBM, ...**
  - **speech translation: ASR + MT**
  - **challenge: MT robust wrt ASR errors $\rightarrow$ data-driven methods**
  - **first research prototype for unlimited domain**
    - **fully automatic, not real time**
    - **without deep learning!**
- **GALE 2005-2010: funded by US DARPA**
- **BOLT 2011-2016: funded by US DARPA**
- **QUAERO 2008-2013: funded by OSEO France**
- **BABEL 2012-2016: funded by US IARPA**
- **EU projects 2011-2015: EU-Bridge, TransLectures**
- **EU ERC advanced grant 2017-2021**
- **Google focused research award 2018-2020**

HLT

ASR: first research 1975-1980

ASR is sequence-to-sequence processing:

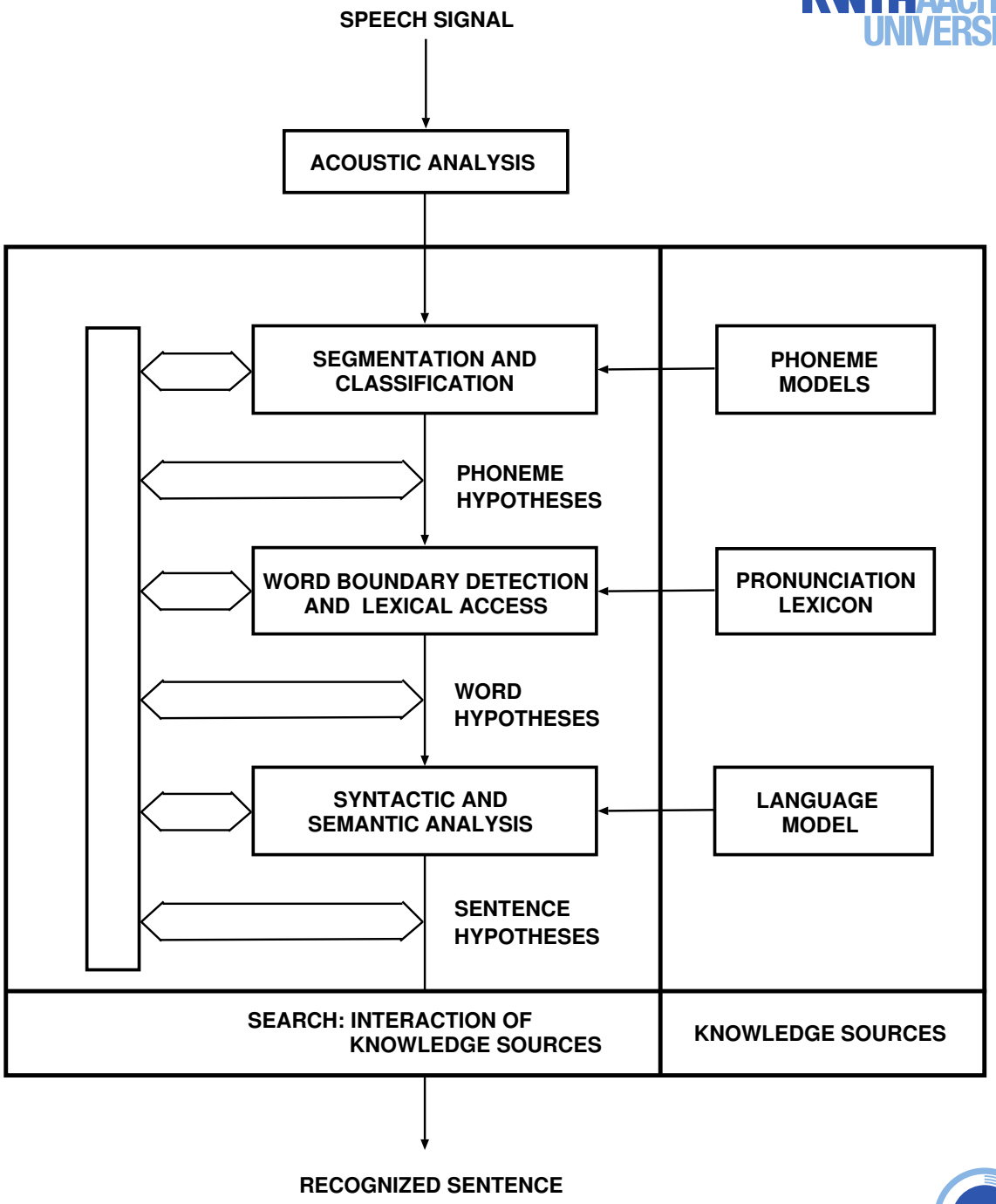
- **sequence of 10-ms acoustic vectors**
- **sequence of sounds/phonemes**
- **sequence of letters**
- **sequence of words**

problems:

- **ambiguities at all levels**
- **interdependencies of decisions**

approach 1975-1980 (Baker/CMU and Jelinek/IBM):

- **score hypotheses using probabilitistic modelling**
- **holistic approach: Bayes decision rule**



principles [textbook by Duda & Hart 1973, p. 11-16]:

- **two strings: observed input x and generated output $c = \hat{c}(x)$ with a (true) probabilistic dependence $pr(c|x)$,**

  from theory to practice: empirical distribution $pr(c|x)$ based on a conceptually "huge" set of string pairs $(c_r, x_r), r = 1, ..., R$

- **performance measure or loss (error) function: $L[\tilde{c}, c]$ between correct output $\tilde{c}$ and output c generated by the system**

- **general form: Bayes decision rule optimizes expected performance, i. e. posterior expectation of loss function based on input x:**

$$x \rightarrow \hat{c}(x) \;:=\; \arg\min_{c} \left\{ \sum_{\tilde{c}} pr(\tilde{c}|x) \cdot L[\tilde{c}, c] \right\}$$

open issue: how to fill in the implementation details?

two conditions for guaranteed optimality:

– exact implementation of loss function
– the true distribution $pr(c|x)$ is known

Bayes Decision Rule: Does the exact form of the loss function matter?

- general form: Bayes decision rule optimizes expected performance with suitable loss function $L[\tilde{c}, c]$:

$$x \rightarrow \hat{c}(x) \; := \; \arg\min_{c} \left\{ \sum_{\tilde{c}} pr(\tilde{c}|x) \cdot L[\tilde{c}, c] \right\}$$

- used in practice (simplification, MAP rule): rule for minimum sequence error:

$$x \rightarrow \hat{c}(x) := \arg\max_{c} \left\{ pr(c|x) \right\}$$

- mathematical equivalence of the two rules [Schlüter & Scharrenbach$^+$ 05, Schlüter & Nussbaum$^+$ 11]:
  - conditions: a metric loss function and $\max_c pr(c|x) \geq 0.5$
  - theoretical refinements beyond the threshold of 0.5
  - experimental results: hard to find a difference
    e. g. for high error rates: from 41% to 40%
- special case for edit distance: improvements beyond MAP rule by position-dependent symbol posterior probabilities [Xu & Povey$^+$ 10, Schlüter & Nussbaum$^+$ 11]

Optimum Performance in Practice

conceptual problem with Bayes decison theory:
– basic assumption: true distribution is known
– in practice: we replace the true distribution by a (learned) model
– question: does the optimality of Bayes decison rule still hold?

we consider mismatch conditions for the classification error:

- **empirical (=true) distributions $pr(c,x)$ and $pr(c|x)$:**
  $E_* =$ Bayes classification error: absolute optimum
- **probability model $p_\vartheta(c|x)$ with set of parameters ϑ:**
  $E_\vartheta =$ model-based classification error using:

$$x \rightarrow \hat{c}_\vartheta(x) = \text{argmax}_c\{p_\vartheta(c|x)\}$$

upper bound on the squared difference between these errors [Ney 03]
(= Kullback-Leibler divergence or relative entropy):

$$\begin{aligned} 1/2 \cdot \left[E_* - E_\vartheta\right]^2 &\leq \sum_{c,x} pr(c,x) \log \frac{pr(c|x)}{p_\vartheta(c|x)} \\ &= \sum_{c,x} pr(c,x) \log pr(c|x) - \sum_{c,x} pr(c,x) \log p_\vartheta(c|x) \end{aligned}$$

suitable training criterion: **minimize the upper bound**

$$1/2 \cdot \min_\vartheta \left[E_* - E_\vartheta\right]^2 \leq \sum_{c,x} pr(c,x) \log pr(c|x) - \max_\vartheta \left\{ \sum_{c,x} pr(c,x) \log p_\vartheta(c|x) \right\}$$

we have derived the cross-entropy training criterion:

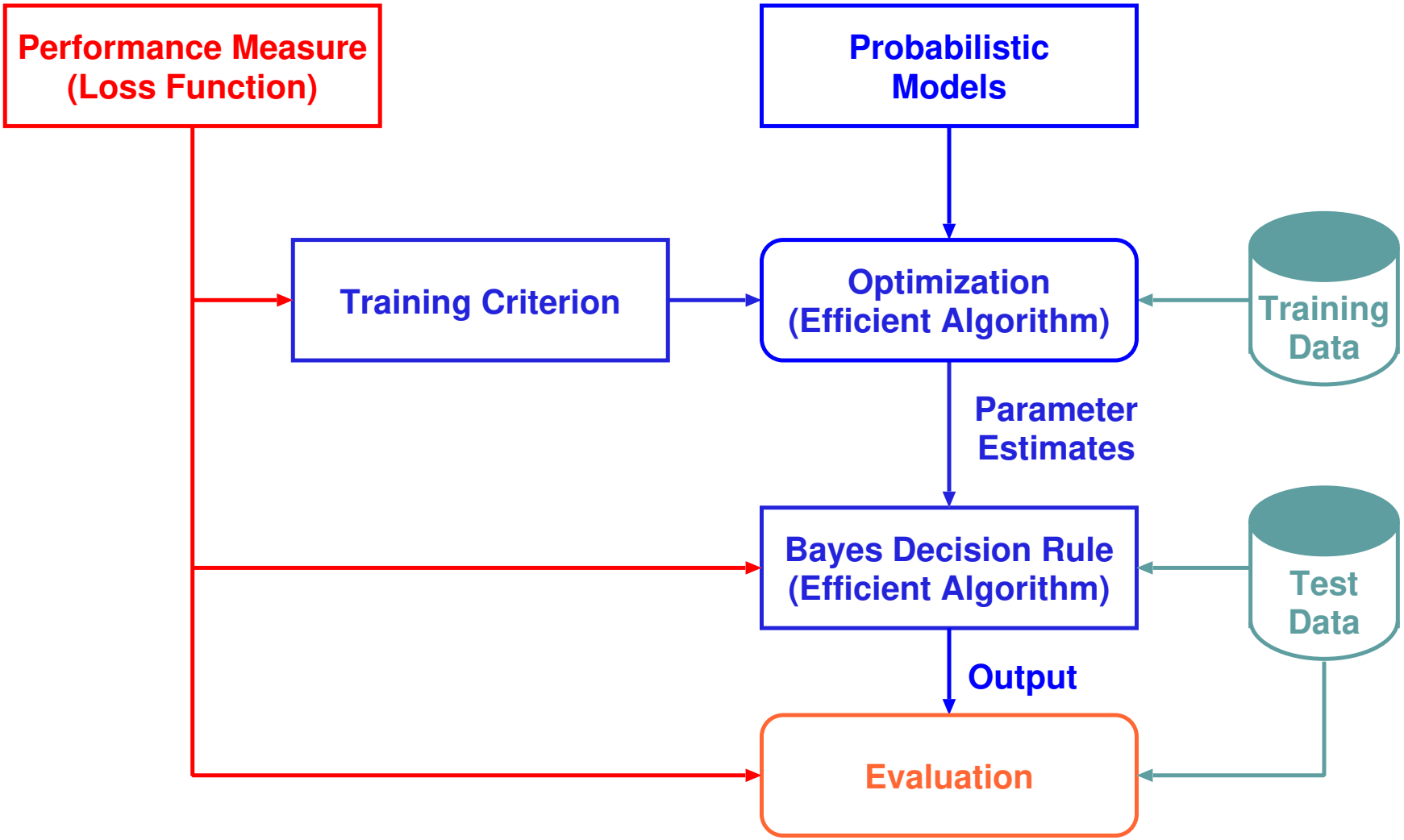
$$\begin{aligned}\hat{\vartheta} &= \operatorname*{argmax}_{\vartheta} \left\{ \sum_{c,x} pr(c,x) \log p_\vartheta(c|x) \right\} \\ &= \operatorname*{argmax}_{\vartheta} \left\{ \sum_{r} \log\, p_\vartheta(c_r|x_r) \right\}\end{aligned}$$

using the "huge" finite set of pairs $(c_r, x_r), r = 1, ..., R$ of the empirical distribution

considerations:

- **resulting training criterion: cross-entropy**
  without proof: well defined optimization problem
- **ASR: various types of outputs and associated errors:**
  – frames without output context: frame-wise cross entropy
  – full sequence context: *sequence discriminative training*
  (ASR: MMI = maximum mutual information)
  · sequence error rate
  · symbol error rate: phones, letters, words
  (details omitted)

Machine Learning: Statistical Decision Theory



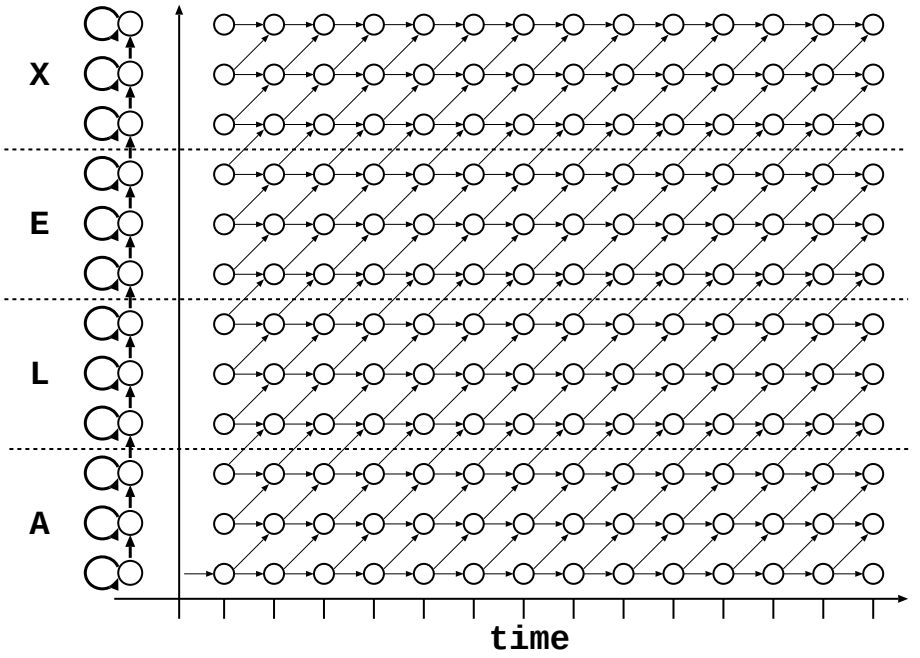
four ingredients:

- performance measure, error measure, cost function:
  – how to judge the quality of the system output
  – examples: ASR: edit distance; MT: TER or BLEU
- probabilistic models (with a suitable structure)
  for capturing the dependencies within and between input and output strings:
  – Markov chain, CRF, (LSTM) RNN, ...
  – generative/hybrid HMM, CTC, neural attention models, ...
- training criterion:
  to learn the free model parameters from examples
  – ideally should be linked to performance criterion
  – two open questions: exact form of criterion? optimization strategy?
- Bayes decision rule: decoder/search
  for generating the output word sequence
  – combinatorial problem (efficient algorithms)
  – should exploit structure of models
  examples: dynamic programming and beam search, A* and heuristic search, ...

Acoustic Modelling: Hidden Markov Model (HMM)

– sequence of acoustic vectors:
$X = x_1^T = x_1 ... x_t ... x_T$ over time t

– sequence of states $s = 1, ..., S$
$s_1^T = s_1 ... s_t ... s_T$ over time t
with associated state labels:
$a_1^S = a_1 ... a_s ... a_S$
$= W$: word sequence



objective of HMM: time alignment
= synchronization between input and output

- classical HMM: generative model for x_1^T:

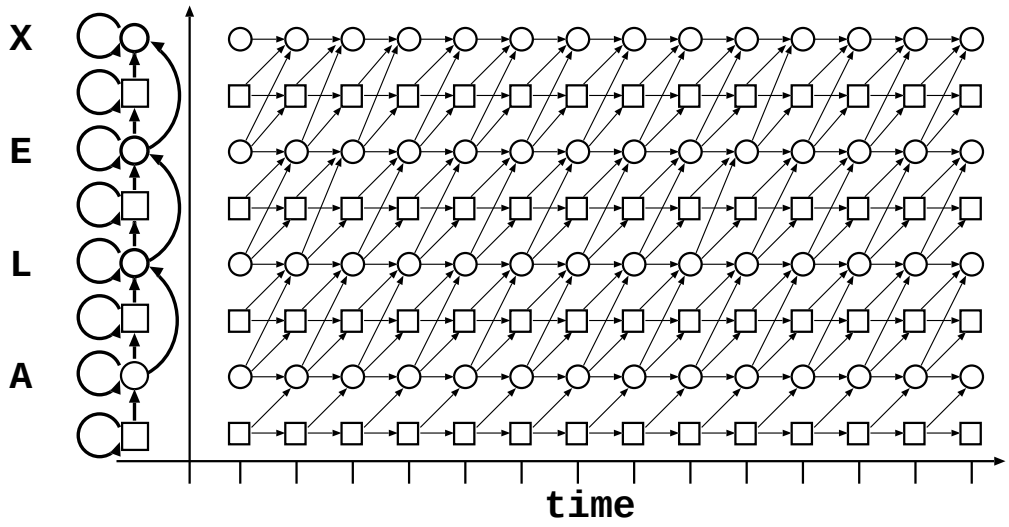
$$q_\vartheta(x_1^T | W = a_1^S) = \sum_{s_1^T} \prod_t q(s_t | s_{t-1}, W) \cdot q_t(x_t | a_{s=s_t})$$

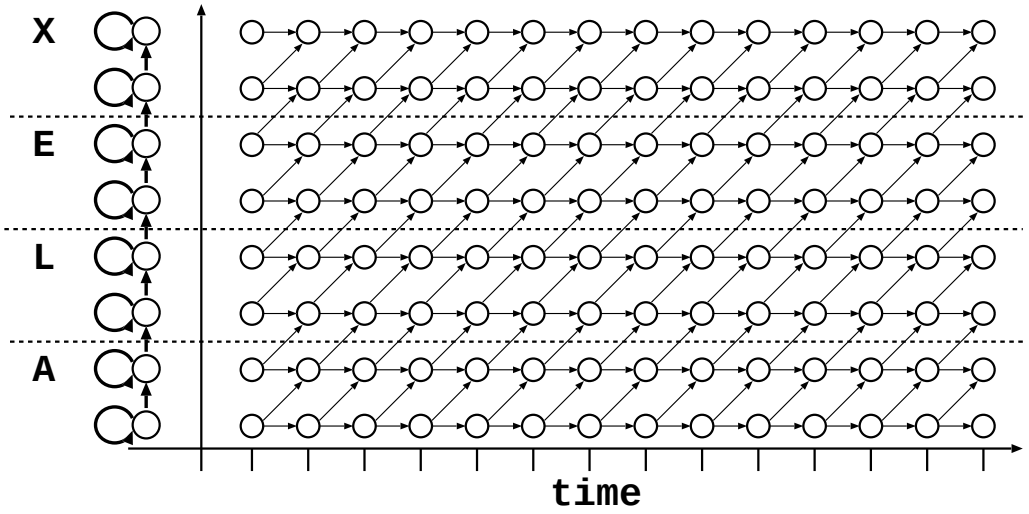
- hybrid HMM: model of label posterior sequence a_1^S:

$$q_\vartheta(W = a_1^S; x_1^T) = \sum_{s_1^T} \prod_t q(s_t | s_{t-1}, W) \cdot q_t(a_{s=s_t} | x_1^T)$$

justification: it is easier to model $q_t(a_s | x_1^T)$ than $q_t(x_t | a_s)$
[Bourlard & Wellekens 89], CTC: [Graves & Fernandez$^+$ 06]

Comparison: CTC vs. Two-State Hybrid HMM





CTC: hybrid HMM with special structure

– no transition probabilities

– blank state (as a separator)

– underlying ANN: LSTM RNN (most important!)

sequence posterior probability $p_\vartheta(W|X)$ with output W and input X using either a (strict) generative or log-linear model:

$$p_\vartheta(W|X) = \frac{q(W) \cdot q_\vartheta(X|W)}{\sum_{\tilde{W}} q(\tilde{W}) \cdot q_\vartheta(X|\tilde{W})} \qquad p_\vartheta(W|X) = \frac{q^\alpha(W) \cdot q_\vartheta^\beta(W;X)}{\sum_{\tilde{W}} q^\alpha(\tilde{W}) \cdot q_\vartheta^\beta(\tilde{W};X)}$$

with the two basic model components:

- **language model (LM) $q(W)$:**
  - **prior that scores the syntactic-semantic adequacy of a sequence**
  - **idea: can be learned from text data only (e. g. 100M words)**
  - **no manual annotation required!**
- **acoustic model (AM): HMM**
  learned from manually transcribed audio data (e. g. 300 hrs = 3M words)
  - **generative: $q_\vartheta(X|W)$ using Gaussian mixtures**
  - **hybrid: $q_\vartheta(W;X)$ using ANN outputs**

ASR: Training Criterion

sequence posterior probability:

$$p_\vartheta(W|X) = \frac{q(W) \cdot q_\vartheta(X|W)}{\sum_{\tilde{W}} q(\tilde{W}) \cdot q_\vartheta(X|\tilde{W})} \qquad p_\vartheta(W|X) = \frac{q^\alpha(W) \cdot q_\vartheta^\beta(W;X)}{\sum_{\tilde{W}} q^\alpha(\tilde{W}) \cdot q_\vartheta^\beta(\tilde{W};X)}$$

suitable training criterion for (audio,text) pairs $(X_r, W_r),\ r = 1, ..., R$ **:**

$$\max_\vartheta \left\{ \sum_r \log\ p_\vartheta(W_r|X_r) \right\}$$

cross-entropy criterion: hard optimization problem due to denominator

- **baseline training: ignore denominator**
  – generative HMM: max. likelihood using EM algorithm (or Viterbi approximation)
  – hybrid HMM: cross-entropy (sum or best path) using backpropagation
  result: no effect of LM on training AM!
- **complete criterion: *sequence discriminative training***
  better approximation: approximate sum in denominator
  – use word hypothesis lattice
  – use simplifed language model (e. g. phoneme fourgram LM)
  result: LM affects training of AM!
- **history (mathematical optimization problem!):**
  Bahl et al./IBM 1986, Normandin 1991, Valtchev 1996,
  Povey 2002 and 2016, Heigold 2005 and 2012

ASR: Sequence Discriminative Training and *End-to-End* Concept

reconsider training criterion for (audio,text) pairs $(X_r, W_r),\ r = 1, ..., R$:

$$\max_{\vartheta} \left\{ \sum_r \log\ p_\vartheta(W_r|X_r) \right\} \qquad p_\vartheta(W|X) := \frac{q^\alpha(W) \cdot q_\vartheta^\beta(W;X)}{\sum_{\tilde{W}} q^\alpha(\tilde{W}) \cdot q_\vartheta^\beta(\tilde{W};X)}$$

different variants of *sequence discriminative training* in ASR:
– sequence error (see above)
– symbol error in sequence context: output labels or frame labels
related concepts: Povey's minimum word/phoneme error rate,
state-level minimum Bayes risk (sMBR)

terminology: What does *end-to-end* mean?

- **training criterion: one global criterion for optimum performance, independent of model structure**
- **optimization strategy of the criterion: one monolithic strategy? e. g. gradient search and backpropagation**
- **monolithic structure of a model:**
  – maybe simplicity/elegance of programming?
  – what about adequacy/performance?

History: ANN in Acoustic Modelling

ANN approaches in ASR:

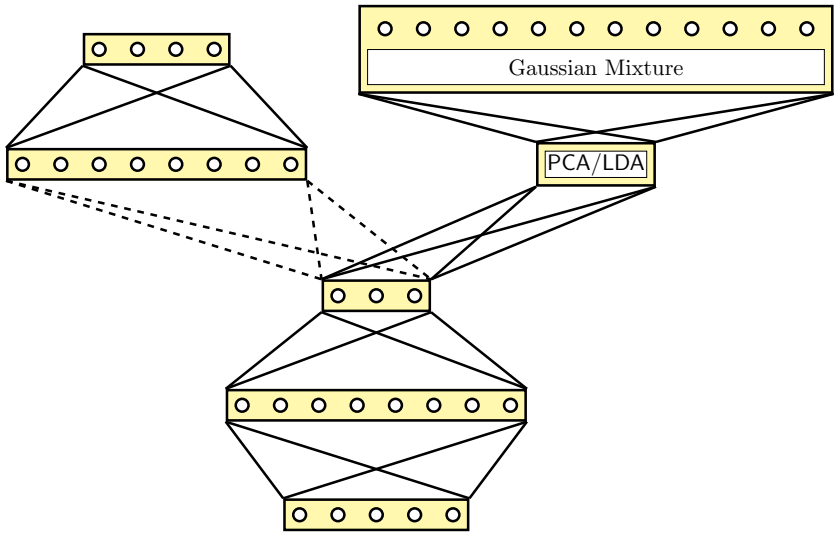
- **1988 [Waibel & Hanazawa+ 88]:**
  phoneme recognition using time-delay neural networks (convolutional NNs!)
- **1989 [Bridle 89]:**
  softmax operation ('Gaussian posterior') for normalization of ANN outputs
- **1989 [Bourlard & Wellekens 89]:**
  – for squared error criterion, ANN outputs can be interpreted as class posterior probabilities (rediscovered: [Patterson & Womack 66])
  – hybrid HMM:
  we replace the emission probabilities by ANN ouputs
- **1993 [Haffner 93]: sum over label-sequence posterior probabilities in hybrid HMMs (CTC like approach)**
- **1994 [Robinson 94]: recurrent neural network in hybrid HMM**
  – competitive results on WSJ task
  – his work remained a singularity in ASR
- **...**

hybrid HMMs until 2008:

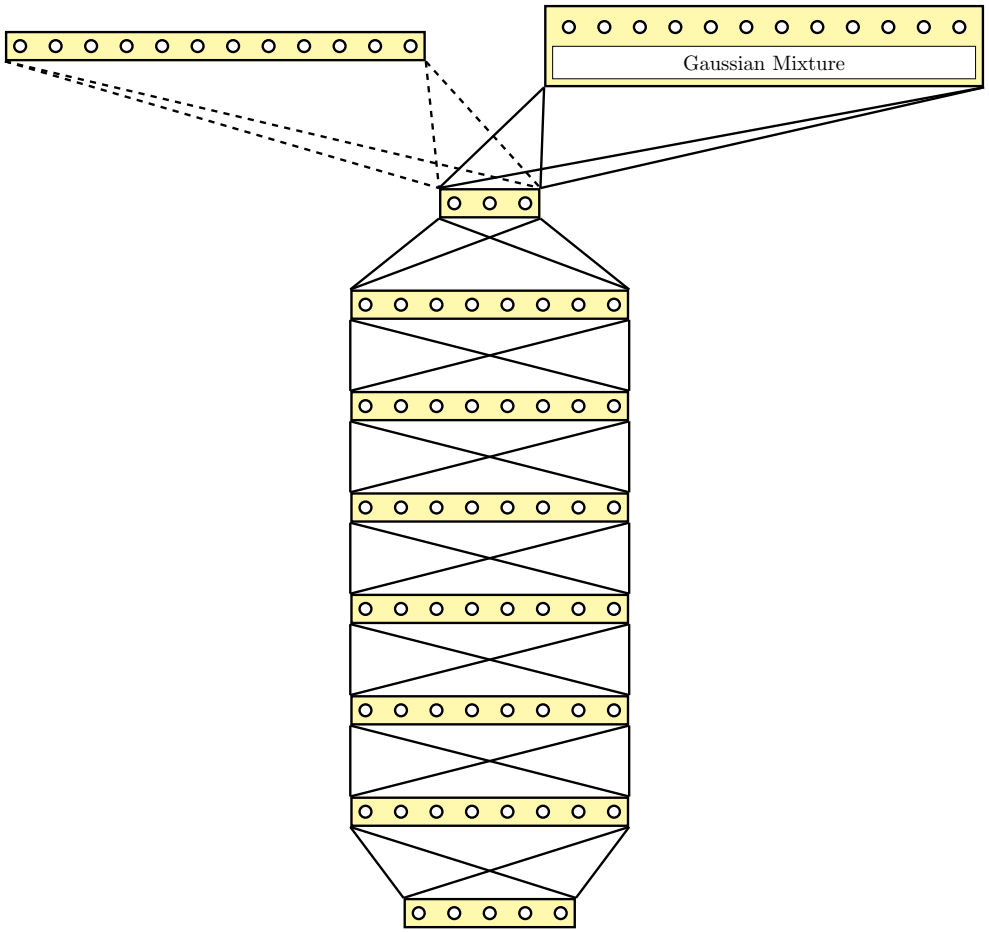
ANNs were never superior to Gaussian mixture models

approach:

- **tandem: use MLP for feature extraction in a generative HMM [Fontaine & Ris$^+$ 97], [Hermansky & Ellis$^+$ 00]**
- **extensions, e. g. bottleneck concept [Stolcke & Grezl$^+$ 06, Grezl & Fousek 08] [Valente & Vepa$^+$ 07, Tüske & Plahl$^+$ 11]**



RWTH's Tandem Structure [Tüske & Plahl$^+$ 11]



History ASR: Tandem vs. Hybrid Approach

RWTH AACHEN UNIVERSITY

tandem approaches:

- **2000 [Hermansky & Ellis$^+$ 00]: multiple layers of processing by combining Gaussian model and ANN for ASR**
- **2006 [Stolcke & Grezl$^+$ 06]: cross-domain and cross-language portability**
- **2007 [Valente & Vepa$^+$ 07]: 8% WER reduction on LVCSR**
- **2011 [Tüske & Plahl$^+$ 11]: 22% WER reduction on LVCSR**

hybrid approaches:

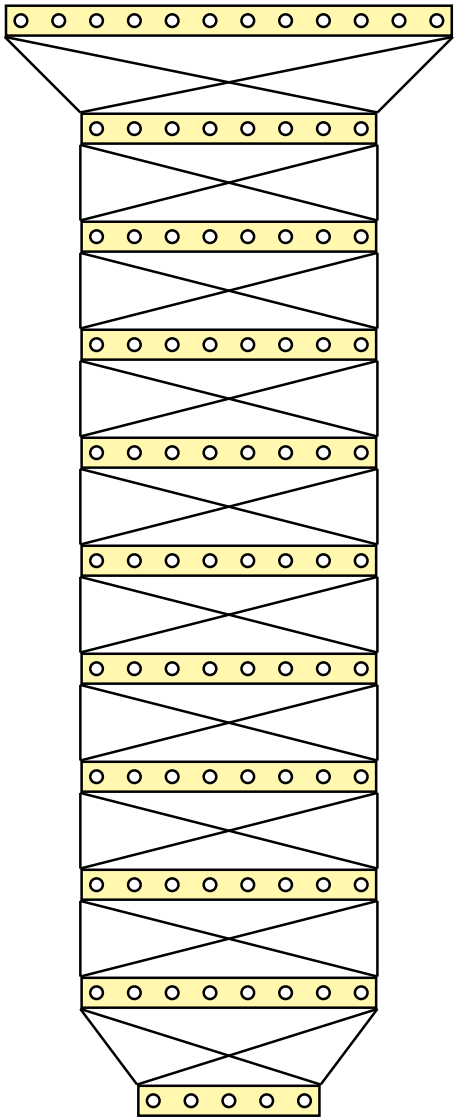
- **2008 [Graves 08]: good results on LSTM RNN for handwriting task (in CTC context)**
- **2010 [Dahl & Ranzato$^+$ 10]: improvement in phone recognition on TIMIT**
- **2011 [Seide & Li$^+$ 11, Dahl & Yu$^+$ 12]: Microsoft Research**
  - **– fully-fledged hybrid approach**
  - **– 30% WER reduction on Switchboard 300h**
- **since 2012: other teams confirmed reductions of WER by 20% to 30%**

comparison: hybrid vs. tandem approach:

- **– hybrid approach: more monolithic and compact**
- **– the same structure in training and testing**
- **– widely used nowadays**

HLT

Artificial Neural Networks (ANN) and Deep Learning:

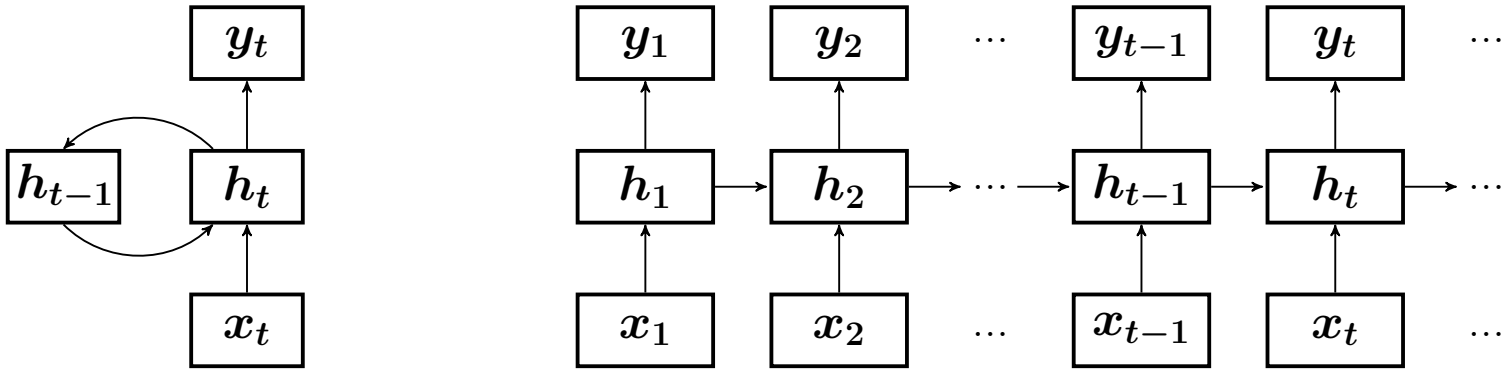


question: what is different now after 30 years?

answer: we have learned how to (better) handle a complex mathematical optimization problem:

- **more powerful hardware (e. g. GPUs)**
- **empirical recipies for optimization: practical experience and heuristics, e.g. layer-by-layer pretraining**
- **result: we are able to handle more complex architectures (deep MLP, RNN, etc.)**

ASR: sequence-to-sequence processing



from simple ANN to RNN:
– introduce a memory (or context) component to keep track of history
– result: two types of input at time t: memory h_{t-1} and observation x_t

extensions:

- **bidirectional structure [Schuster & Paliwal 97]**
- **LSTM: long short-term memory**
  [Hochreiter & Schmidhuber 97, Gers & Schraudolph$^+$ 02]

Systematic Experiments on QUAERO English Eval 2013 (Tueske et al. RWTH 2017)

RWTH AACHEN UNIVERSITY

QUAERO task: broadcast news/conversations, podcasts, TED lectures

Word error rates [%] on QUAERO English Eval 2013
(note: acoustic input features were optimized for acoustic model):

Acoustic Model			Language Model	
Model		Criterion	Count	Count+ANN
			PP=131.1	PP=92.0
Gaussian Mixtures		Max.Lik.	20.7	
		seq.disc. training	19.2	16.1
Neural Net	FF MLP	frame-wise CE	11.6	
		seq.disc. training	10.7	9.0
	LSTM RNN	frame-wise CE	10.6	
		seq.disc. training	9.8	8.2

observations:
– improvements by acoustic ANNs: 50% relative
– improvement by language model ANN: 15% relative
– total improvements: 60% relative

HLT

Tasks: Switchboard and Call Home

- conversational speech: telephone speech, narrow band;
  challenging task: initial WER: 60% on Switchboard

- training data for acoustic model: Switchboard corpus
  - about 300 hours of speech
  - about 2400 two-sided recordings with an average of 200 seconds
  - 543 speakers
- test set Hub5'00
  - 20 telephone recordings form Switchboard studies (SWB)
  - 20 telephone conversations from Call-Home US English Speech
  - total: 3.5 hours of speech
- training data for language model
  - vocabulary size fixed to 30k
  - Switchboard corpus: 2.9M running words
  - Fisher corpus: 21M running words

ASR: Recent Results (April 2019)

RWTH AACHEN UNIVERSITY

baseline models:
- **language model: 4-gram count model**
- **acoustic model: hybrid HMM with CART (allophonic) labels: LSTM bi-RNN with frame-wise cross-entropy training**
- **speaker/channel adaptation: i-vector [Dehak & Kenny+ 11]**
- **affine transformation [Gemello & Manai+ 06, Miao & Metze 15]**

word error rates [%]:

adaptation	methods	SWB	CHM	average
no	baseline approach	9.7	19.1	14.4
	+ seq. discr. training (sMBR)	9.6	18.3	13.9
	+ LSTM-RNN language model	7.7	15.8	11.7
yes (i-vector)	baseline approach	9.0	18.0	13.5
	+ seq. discr. training (sMBR)	8.4	17.2	12.8
	+ LSTM-RNN language model	6.8	15.1	10.9
+ adaptation by affine transformation		6.7	13.5	10.2

overall improvements over baseline:
- **33% relative reduction in WER**
- **by seq. discr. training, LSTM-RNN language model and adaptation**

HLT

Best Results on Call Home (CHM) and Switchboard (SWB)
(best word error rates [%] reported)

team	CHM	SWB	training data, remarks
Johns Hopkins U 2017	18.1	9.0	300h, no ANN-LM, single model, data perturbation
Microsoft 2017	17.7	8.2	300h, ResNet, with ANN-LM
ITMO U 2016	16.0	7.8	300h, with ANN-LM, model comb., data perturbation
Google 2019/arXiv	14.1	6.8	300h, attention models
RWTH U 2017	15.7	8.2	300h, with ANN-LM, model comb.
RWTH U 2019/arXiv	13.5	6.7	300h, single models, adaptation
Microsoft 2017	12.0	6.2	2000h, model comb.
IBM 2017	10.0	5.5	2000h, model comb.
Capio 2017	9.1	5.0	2000h, model comb.

ASR: Librispeech Task

training data:
– audio : 960 hrs
– text: 800 million words

word error rates[%]:

		dev		test	
team	approach	1st half	2nd half	1st half	2nd half
Irie, Zeyer et al. RWTH 2019	attention with BPE units, no LM	4.3	12.9	4.4	13.5
	+ LSTM-RNN LM	3.0	9.1	3.5	10.0
	+ transformer LM	2.9	8.8	3.1	9.8
Lüscher, Beck et al. RWTH 2019	hybrid HMM, CART, 4g LM	4.3	10.0	4.8	10.7
	+ seq. disc. training	3.7	8.7	4.2	9.3
	+ LSTM-RNN LM	2.4	5.8	2.8	6.2
	+ transformer LM	2.3	5.2	2.7	5.7
Zeghidour et al., FB 2018	gated CNN with letters/words	3.2	10.1	3.4	11.2
Irie et al., Google 2019	attention with WPM units	3.3	10.3	3.6	10.3
Park et al., Google 2019	attention ... data augmentation	-	-	2.5	5.8

goal of language modelling: compute the prior $p(w_1^N)$ of a word sequence w_1^N
– how plausible is this word sequence w_1^N (independently of observation x_1^T!) ?
– measure of language model quality: perplexity PP

$$\log PP := \log 1 \Big/ \sqrt[N]{p(w_1^N)} = -1/N \cdot \sum_{n=1}^{N} \log p(w_n|w_0^{n-1})$$

interpretation: effective vocabulary size as seen by ASR decoder/search

interpretation: prediction task:
based on history w_0^{n-1}, predict $p(w_n|...)$

approaches:
– use full history: RNN or LSTM
– truncate history: $\rightarrow$ k-gram MLP

perplexity PP on test data (QUAERO)
(Sundermeyer et al.; RWTH 2012, 2015):

approach	PP
baseline: count model	163.7
10-gram MLP	136.5
RNN	125.2
LSTM-RNN	107.8
10-gram MLP with 2 layers	130.9
LSTM-RNN with 2 layers	100.5

important result: improvement of PP by 40%

Interpolated Language Models: Perplexity and WER

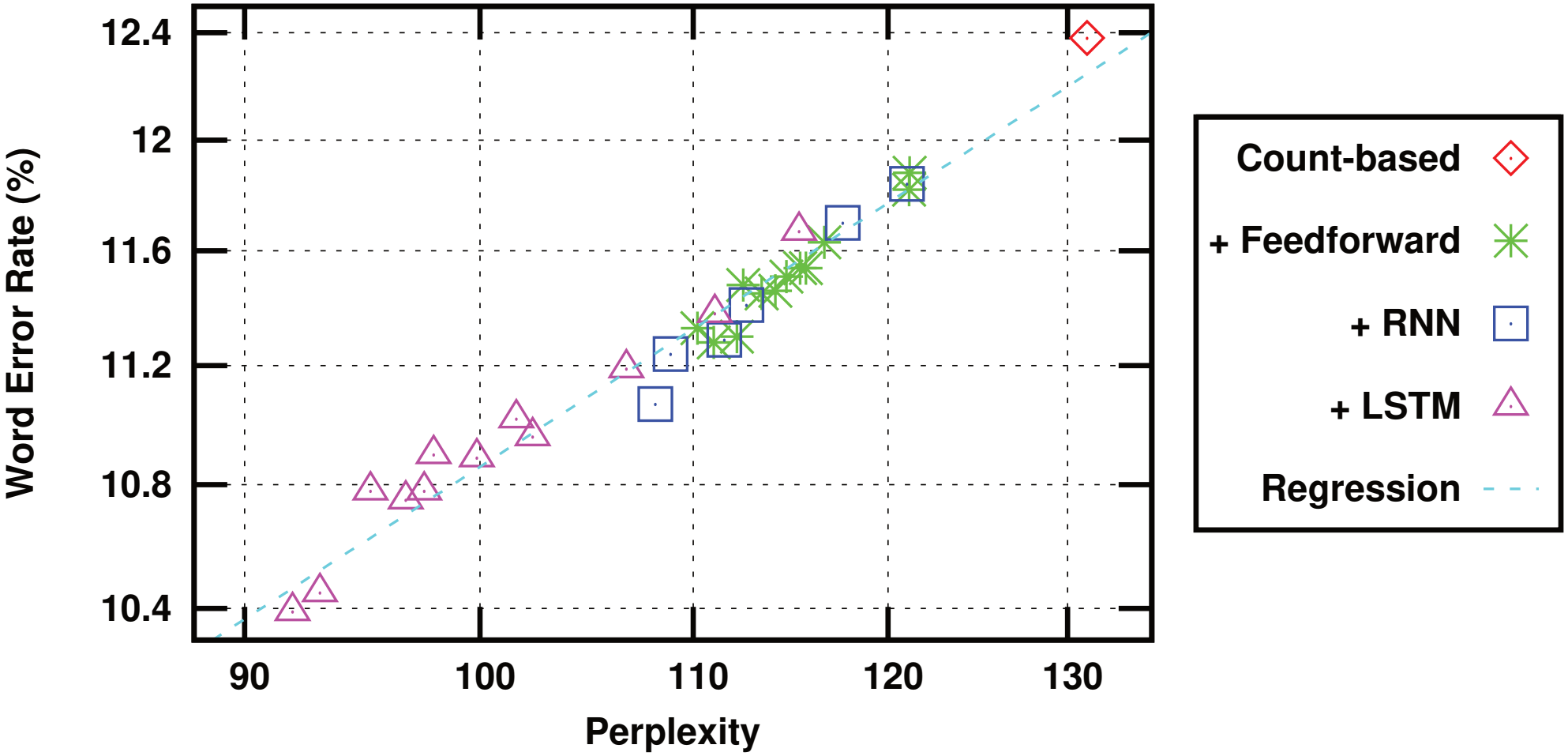
- more details and refinements:
  - use of word classes for softmax in output layer
  - unlimited history of RNN: requires re-design of ASR search
- in practice:
  interpolation of TWO models: count model (3B words) + ANN model (60M words)
- perplexity and word error rate on test data

models	PP	WER[%]
count model	131.2	12.4
+ 10-gram MLP	112.5	11.5
+ Recurrent NN	108.1	11.1
+ LSTM-RNN	96.7	10.8
+ 10-gram MLP with 2 layers	110.2	11.3
+ LSTM-RNN with 2 layers	92.0	10.4

- improvements achieved:
  - perplexity: 30% reduction: from 131 to 92
  - WER: 15% reduction: from 12.4% to 10.4%

Plot: Perplexity vs. Word Error Rate

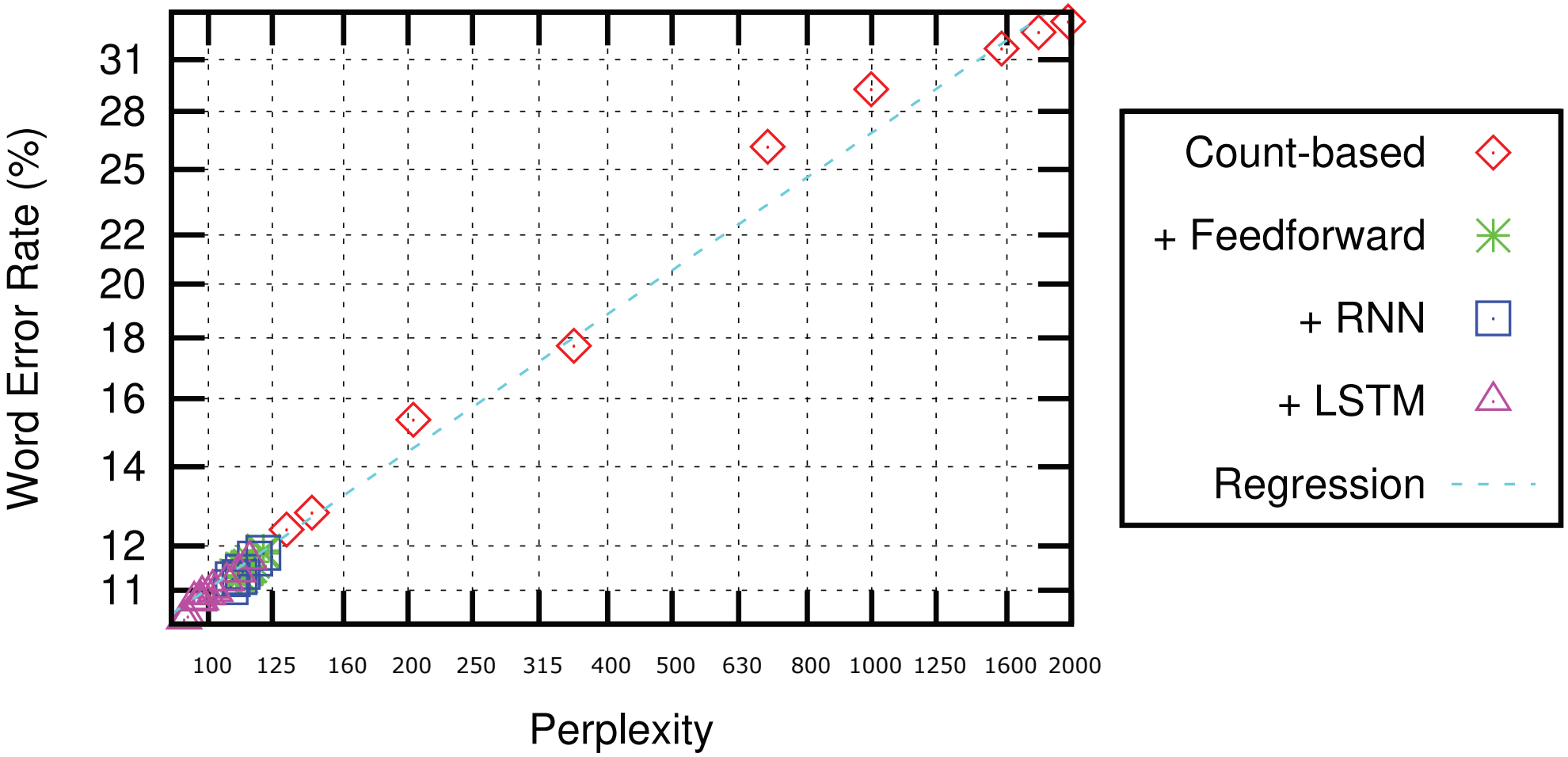
empirical law: $WER = \alpha \cdot PP^{\beta}$
[Makhoul & Schwartz 94, Klakow & Peters 02]



Extended Range: Perplexity vs. Word Error Rate

empirical law: $WER = \alpha \cdot PP^{\beta}$

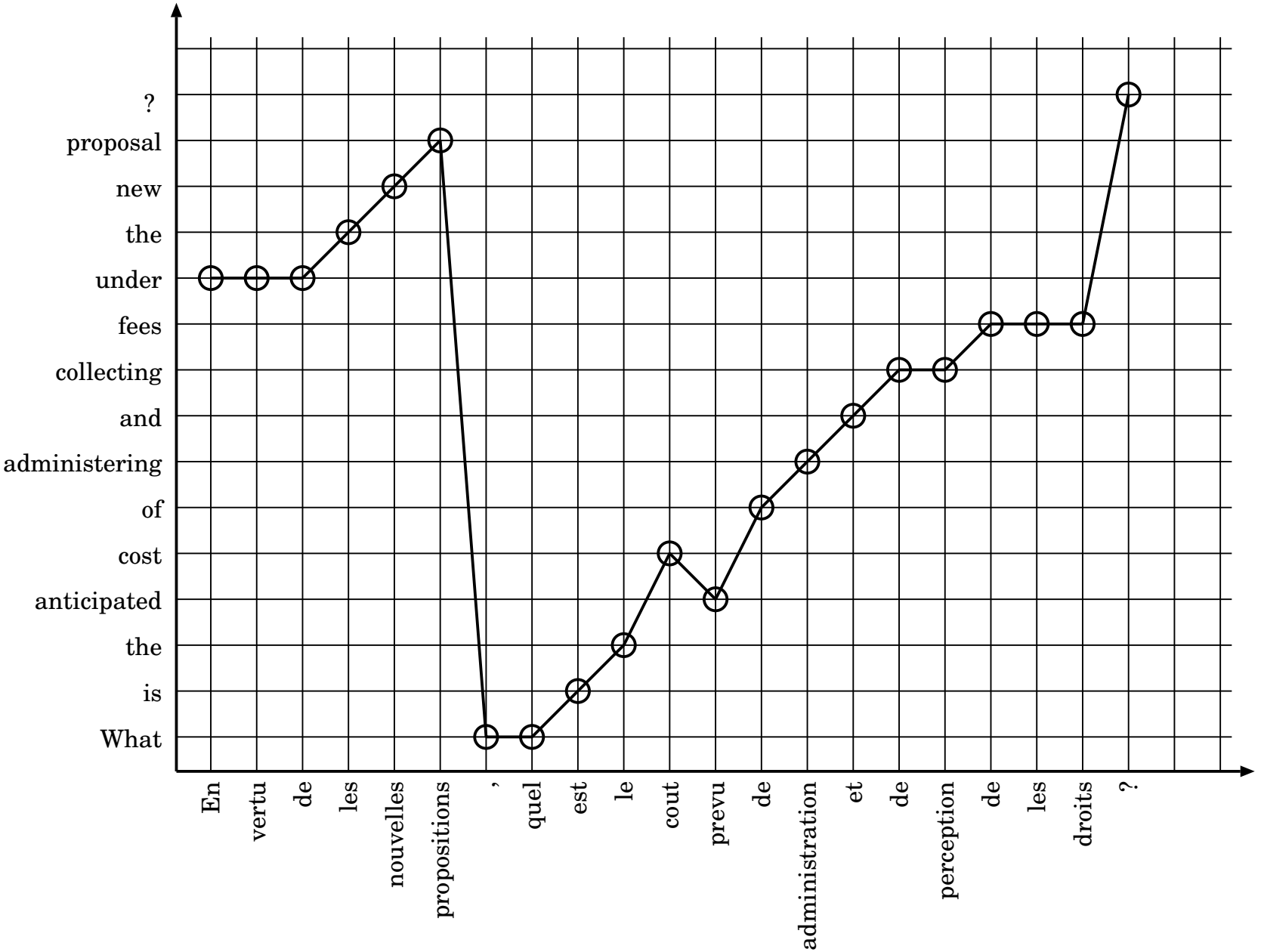
open question: theoretical justification?



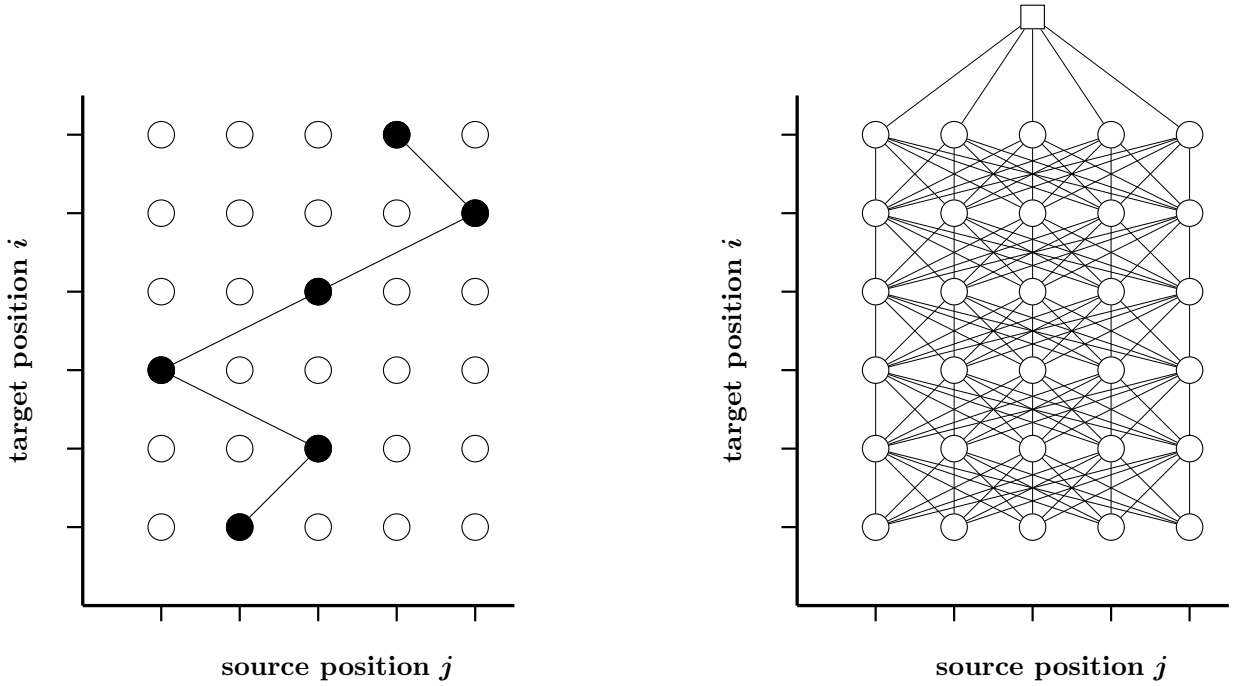
statistical approaches were controversial in MT (and other NLP tasks):

- **1969 Chomsky:**
  ... the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term.
- **until 2000: mainstream approach was rule-based**
  – result: huge human effort required in practice
  – problems: coverage and consistency of rules
- **1989-93: IBM Research: statistical approach to MT**
  1994: key people (Mercer, Brown) left for a hedge fund
- **1996-2002 RWTH: improvements beyond IBM's approach:**
  phrase-based approach and log-linear modelling
- **around 2004: from singularity to mainstream in MT**
  F. Och (and more RWTH PhD students) joined Google
  2008: service *Google Translate*
- **2015: neural MT: attention mechanism [Bahdanau & Cho$^+$ 15]**

Word Alignments
(learned automatically; Canadian Parliament)



Machine Translation: Direct HMM



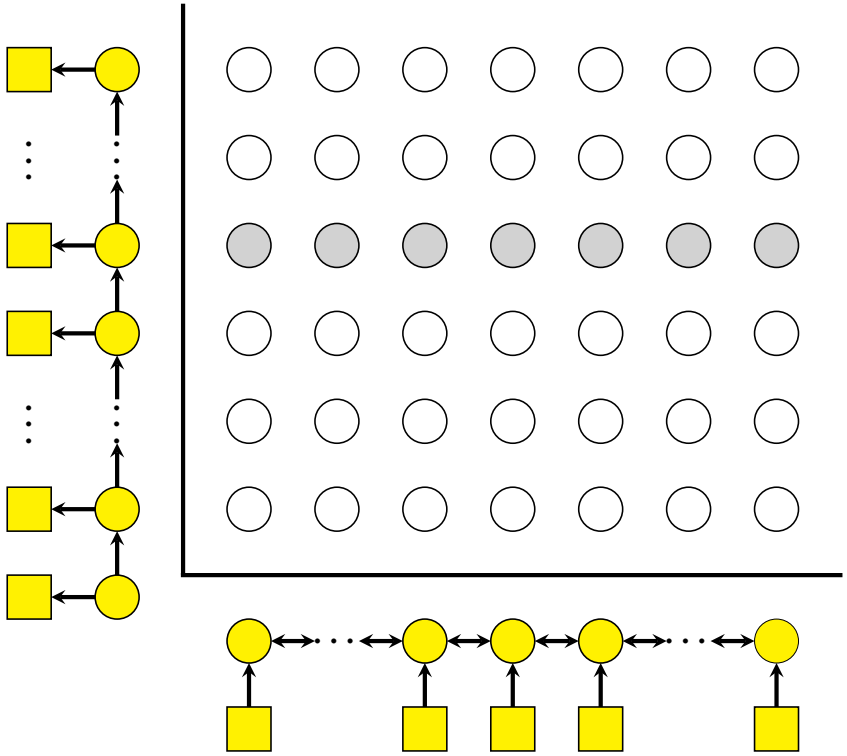
- translation: from source sentence $f_1^J = f_1 ... f_j ... f_J$ to target sentence $e_1^I = e_1 ... e_i ... e_I$
- alignment direction: from target to source: $i \rightarrow j = b_i$
- first-order hidden alignments and factorization:

$$p(e_1^I|f_1^J) = \sum_{b_1^I} p(b_1^I, e_1^I|f_1^J) = \sum_{b_1^I} \prod_i p(b_i, e_i|b_{i-1}, e_0^{i-1}, f_1^J)$$

- resulting model: exploit first-order structure (or zero-order)
  training: backpropagation within EM algorithm

common properties in both approaches:

– bi-directional LSTM RNN over input words f_j, $j = 1, ..., J$
– uni-directional LSTM RNN over output words e_i, $i = 1, ..., I$
– link between input and output: alignment/attention probabilities $p(j|i, e_0^{i-1}, f_1^J)$



use of alignment/attention weights $p(j|i, e_0^{i-1}, f_1^J)$, i. e. distribution over source positions j:

- **attention mechanism: averaging over internal RNN representations**
- **zero-order HMM (mixture model): averaging over probability models**

WMT Task: German → English

WMT task: workshop on machine translation
(bilingual training data: 6M sentence pairs with 160M words in each language)

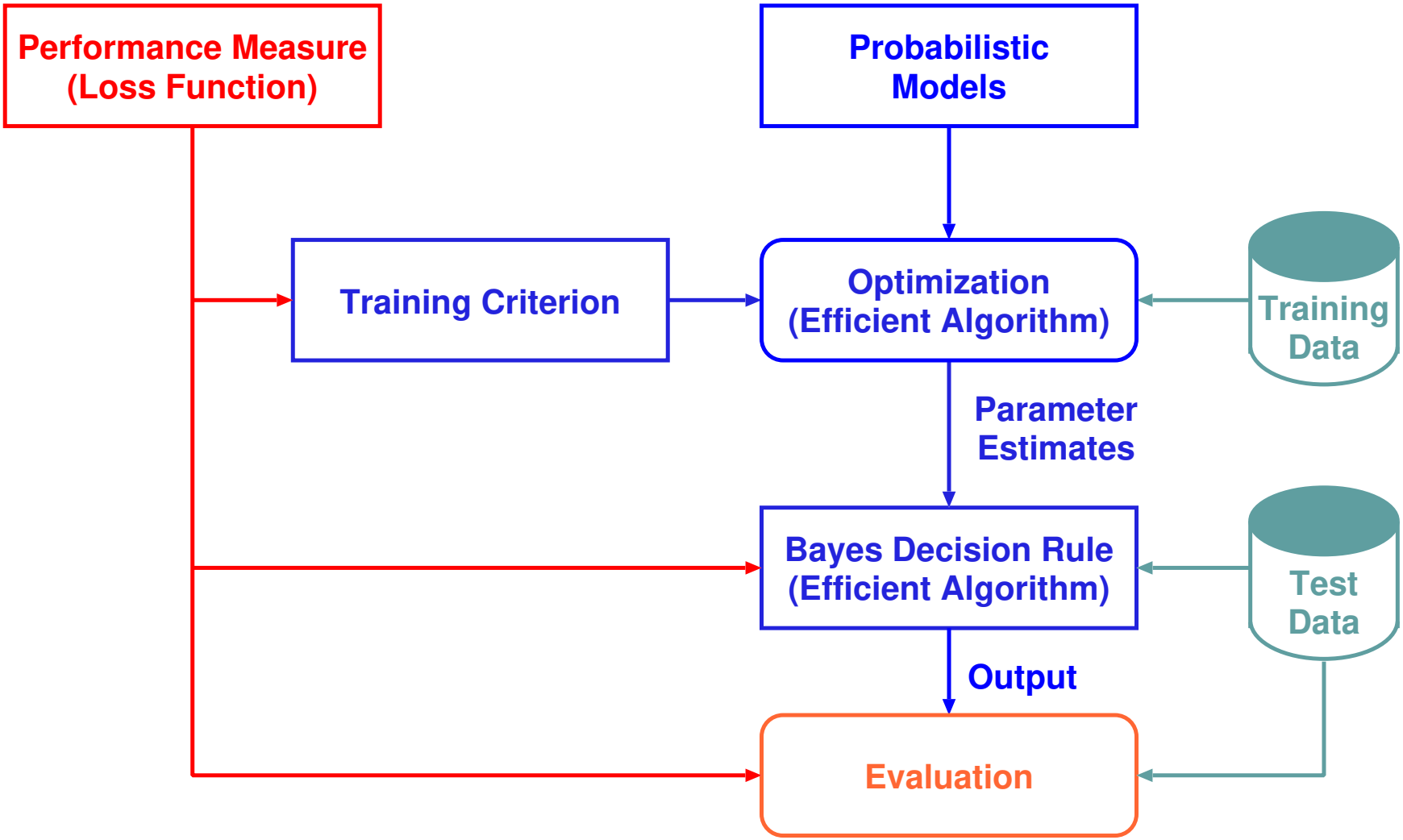
performance measures:
– BLEU [%]: accuracy measure: "the higher, the better"
– TER [%]: error measure: "the lower, the better"

results for various neural MT approaches:
(RWTH: Wang, Bahar; 2018, 2019)

approach	representation	newstest2017 BLEU	newstest2017 TER	newstest2018 BLEU	newstest2018 TER
Direct HMM: 1st-order	RNN 4:1	31.6	56.5	38.7	48.4
0th-order	RNN 4:1	31.6	56.7	38.1	48.9
	self-attention	33.8	54.9	40.5	46.8
Attention	RNN 4:1	32.1	56.3	38.8	48.1
	self-attention	33.4	55.3	40.4	46.8

Summary and Conclusions

- **Bayes decision theory and statistical approach:**
  four key ingredients
  – choice of performance measure: errors at sequence, word, phoneme, frame level
  – probabilistic models at these levels and the interaction between these levels
  – training criterion along with an optimization algorithm
  – Bayes decision rule: search/decoder with an efficient implementation
- **deep learning:**
  – defines one family of probabilistic models within statistical approach
  – baseline structure: matrix-vector product + nonlinearities
  – yes, resulted in significant improvements
- **history of machine learning and statistical classification:**
  there has been and will be life outside deep learning

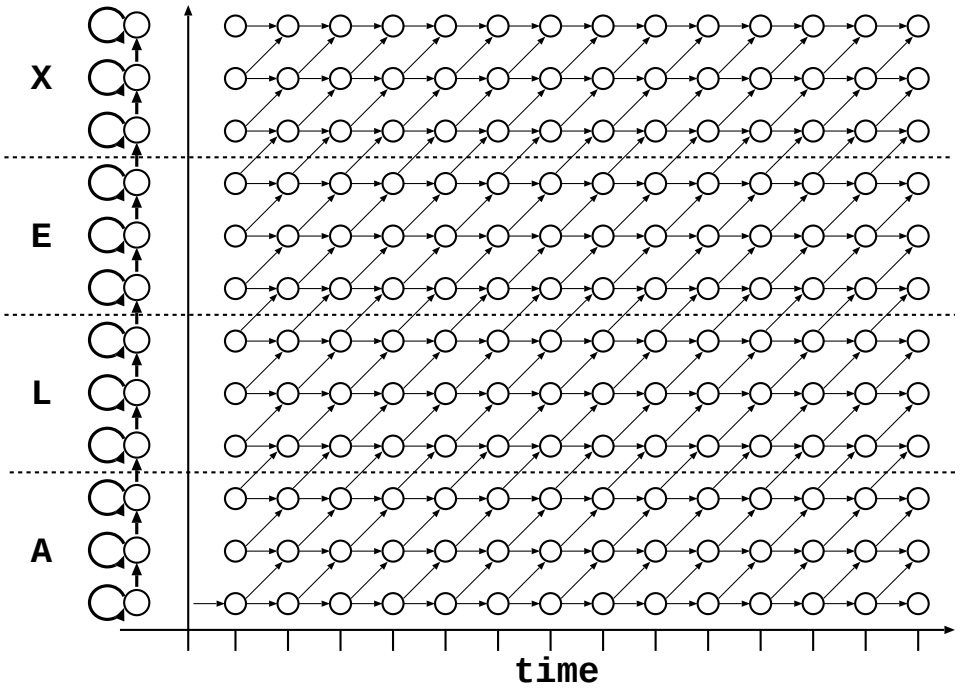

Performance Measure
(Loss Function)
Probabilistic
Models
Training Criterion
Optimization
(Efficient Algorithm)
Training
Data
Parameter
Estimates
Bayes Decision Rule
(Efficient Algorithm)
Test
Data
Output
Evaluation

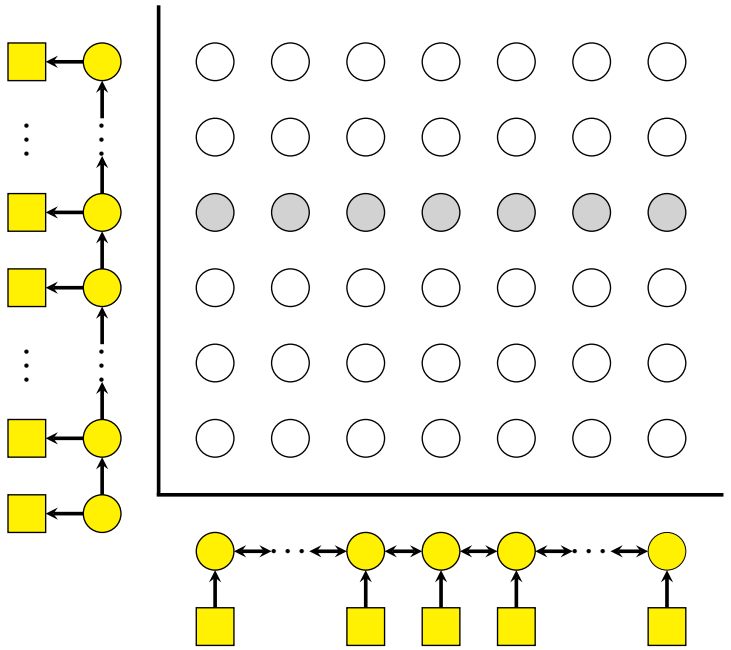
Challenges

- challenges for general machine learning:
  - mathematical optimization with huge complexity:
    we need a theoretical framework for practical aspects of gradient search
  - can we find ANNs with more explicit probabilistic structures?
  - novel structures beyond matrix-vector product + nonlinearties?
- challenges in ASR:
  - to continue the general improvements (ongoing for 40 years!)
  - task: informal colloquial speech (meetings)
  - robustness wrt acoustic conditions and language context
    (improved adaptation ?)
- unsupervised training for ASR:
  machine learning with (virtually) no labeled data?
- features for ASR beyond spectral analysis/Fourier Transform:
  - recent work [Tüske & Golik 2014, Sainath et al. 2015]
  - real improvements: ?

THE END
RAAI Moscow
06-Jul-2019

Sequence-to-Sequence Processing: Direct HMM and Attention Model

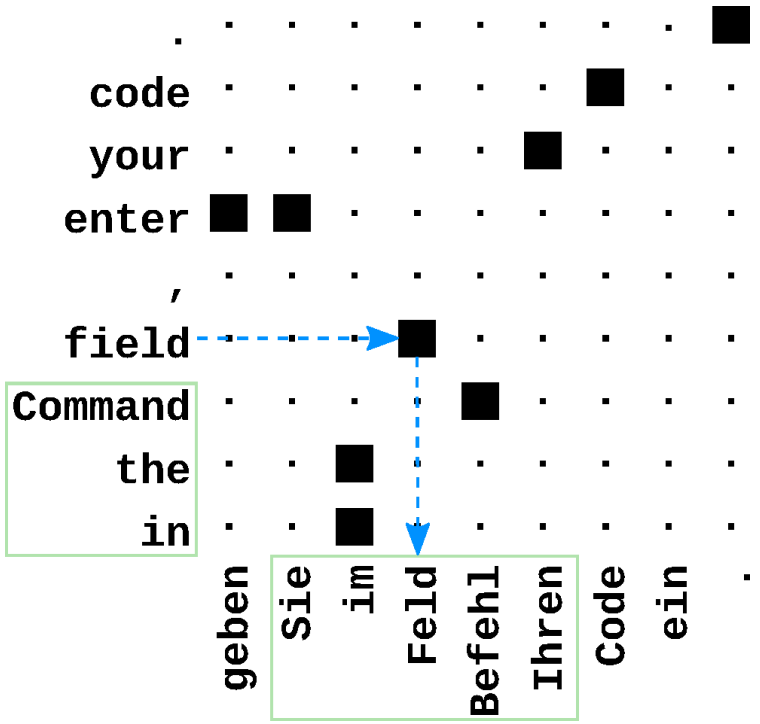




Direct HMM: Dependencies Modelled by FF ANN

decomposition into alignment and lexicon model:

$$p(b_i, e_i | b_{i-1}, e_0^{i-1}, f_1^J) = p(b_i | b_{i-1}, e_0^{i-1}, f_1^J) \cdot p(e_i | b_{i-1}^{i}, e_0^{i-1}, f_1^J)$$



lexicon model:

$$p(e_i | f_{b_i-2}^{b_i+2}, e_{i-3}^{i-1})$$

alignment model:

$$p(\Delta_i | f_{b_{i-1}-2}^{b_{i-1}+2}, e_{i-3}^{i-1})$$

with $\Delta_i := b_i - b_{i-1}$

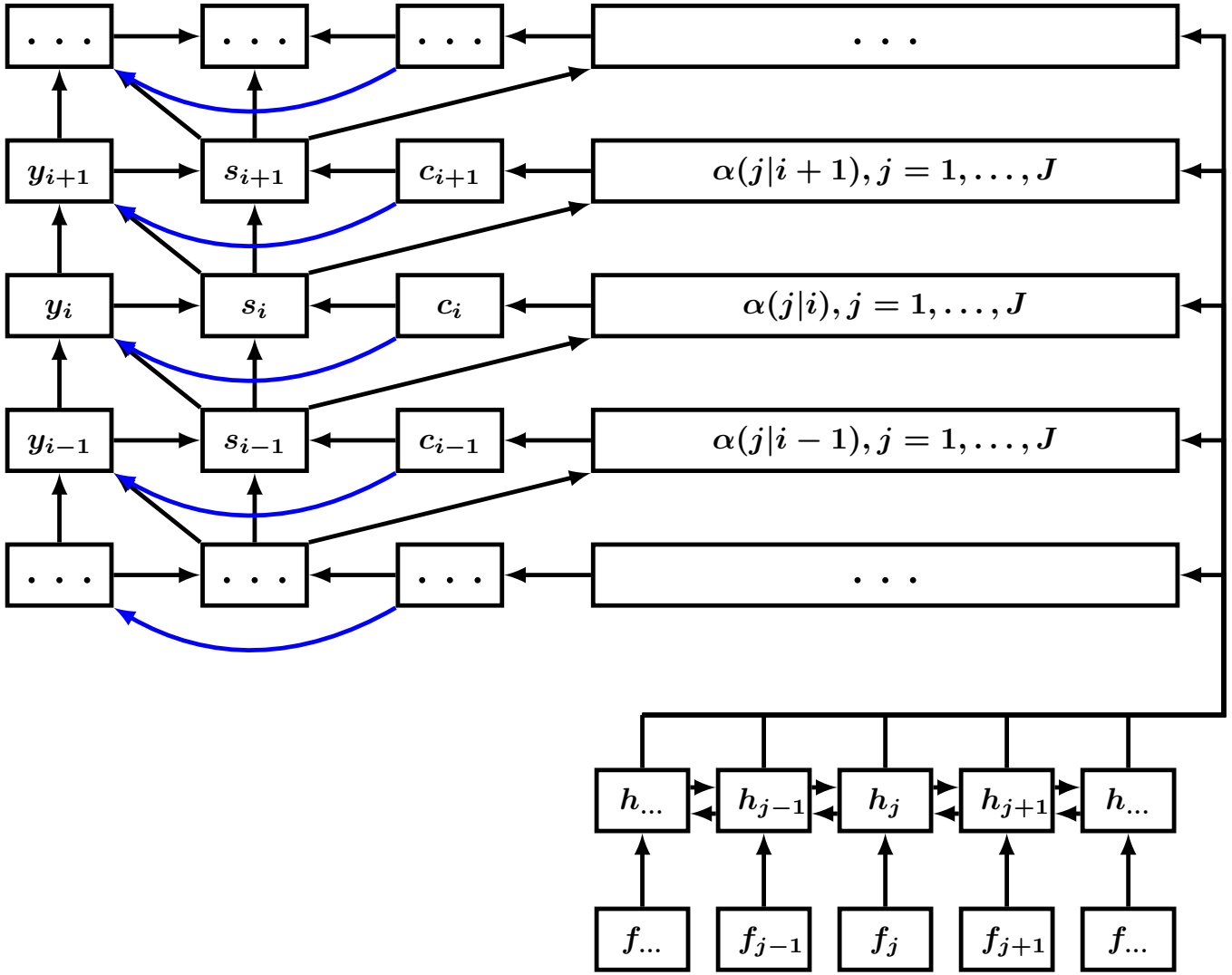
compare with count-based HMM [Vogel & Ney$^+$ 96]:

$$p(b_i, e_i | b_{i-1}, e_0^{i-1}, f_1^J) = p(b_i | b_{i-1}) \cdot p(e_i | f_{b_i})$$

State of the Art: Attention-based Neural MT [Bahdanau & Cho+ 15]

operations:

- **RNN on source side:**
  $f_1^J \rightarrow h_j = H_j(f_1^J)$
- **alignment weights:**
  $\alpha(j|i) = A(s_{i-1}, h_j)$
- **weighted context vector:**
  $c_i = \sum_j \alpha(j|i) \cdot h_j$
- **output vector y_i:**
  $y_i = Y(y_{i-1}, s_{i-1}, c_i)$
  conventional notation:
  $y_i \equiv p_i(e|e_0^{i-1}, h_1^J)$
- **state vector of target RNN:**
  $s_i = S(s_{i-1}, y_i, c_i)$



RNNs with LSTM/GRU extensions on source and target sides

History:

- **1989 [Nakamura & Shikano 89]:**
  English word category prediction based on neural networks.
- **1993 [Castano & Vidal$^+$ 93]:**
  Inference of stochastic regular languages through simple recurrent networks
- **2000 [Bengio & Ducharme$^+$ 00]:**
  A neural probabilistic language model
- **2007 [Schwenk 07]: Continuous space language models**
  2007 [Schwenk & Costa-jussa$^+$ 07]: Smooth bilingual n-gram translation (!)
- **2010 [Mikolov & Karafiat$^+$ 10]:**
  Recurrent neural network based language model
- **2012 RWTH Aachen [Sundermeyer & Schlüter$^+$ 12]:**
  LSTM recurrent neural networks for language modeling

today: ANNs in language show competitive results.

RWTH AACHEN UNIVERSITY

History of NN based approaches to MT:

- **1997 [Neco & Forcada 97]:**
  asynchronous translations with recurrent neural nets
- **1997 [Castano & Casacuberta 97, Castano & Casacuberta$^+$ 97]:**
  machine translation using neural networks and finite-state models
- **2007 [Schwenk & Costa-jussa$^+$ 07]:**
  smooth bilingual n-gram translation
- **2012 [Le & Allauzen$^+$ 12, Schwenk 12]:**
  continuous space translation models with neural networks
- **2014 [Devlin & Zbib$^+$ 14]:**
  fast and robust neural networks for SMT
- **2014 [Sundermeyer & Alkhouli$^+$ 14]:**
  recurrent bi-directional LSTM RNN for SMT
- **2015 [Bahdanau & Cho$^+$ 15]:**
  joint learning to align and translate

HLT

distinguish two approaches to modelling $p(E|F)$:

- traditional approach (in ASR and MT):
  - separate language model $p(E)$ and observation model $p(F|E)$
  - consider posterior probability:

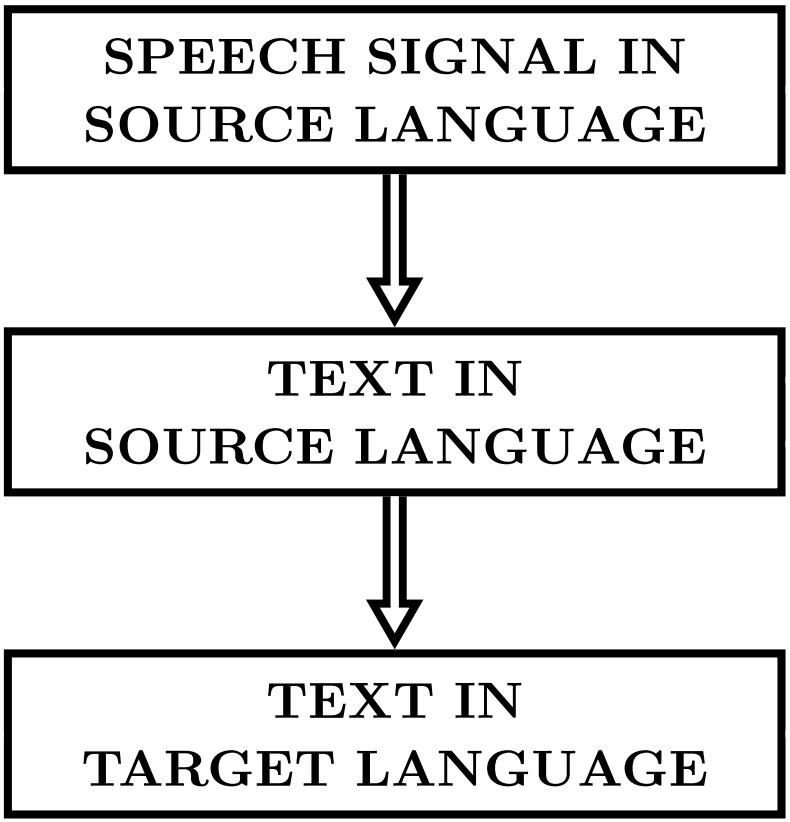
$$p(E|F) = \frac{p(E) \cdot p(F|E)}{\sum_{\tilde{E}} p(\tilde{E}) \cdot p(F|\tilde{E})}$$

  - discriminative approach (MMI in ASR):
    use $p(E|F)$ as training criterion
  - generative approach: ignore denominator in training
    and use maximum likelihood in lieu of MMI
- direct approach (discriminative):
  - start with posterior probability and use factorization for $E = e_1^I = e_1 ... e_i ... e_I$:

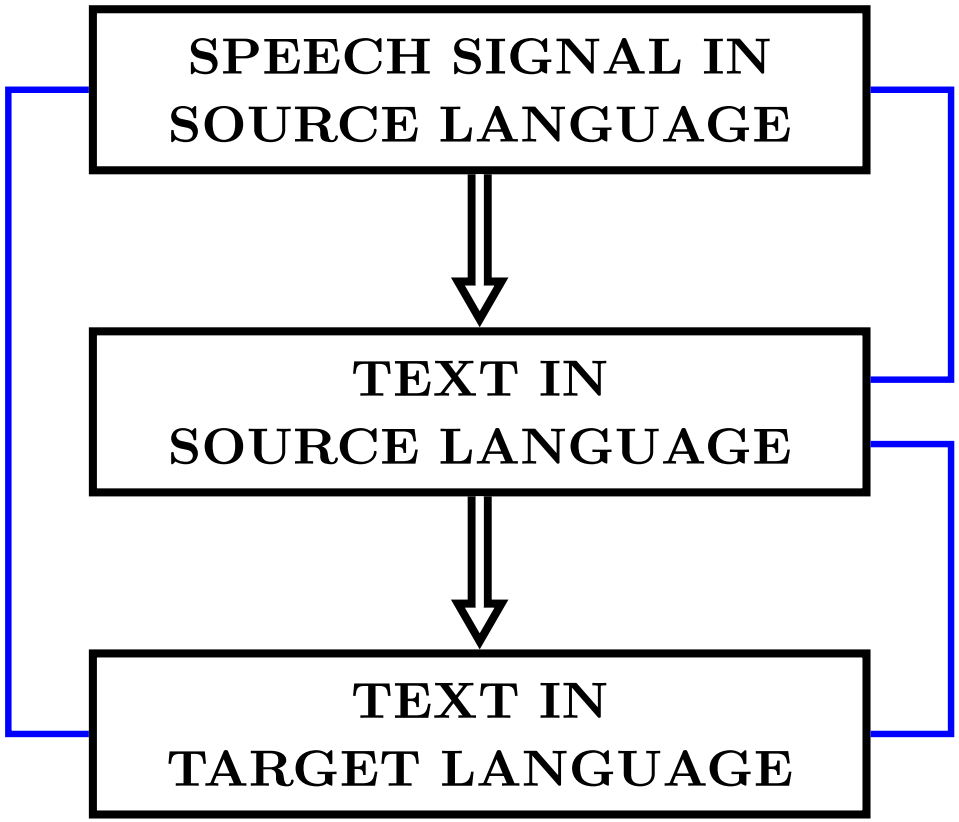
$$p(E|F) = p(e_1^I|F) = \prod_i p(e_i|e_0^{i-1}, F)$$

  - need for a localization mechanism:
    attention model or a alignment model (as in HMM)

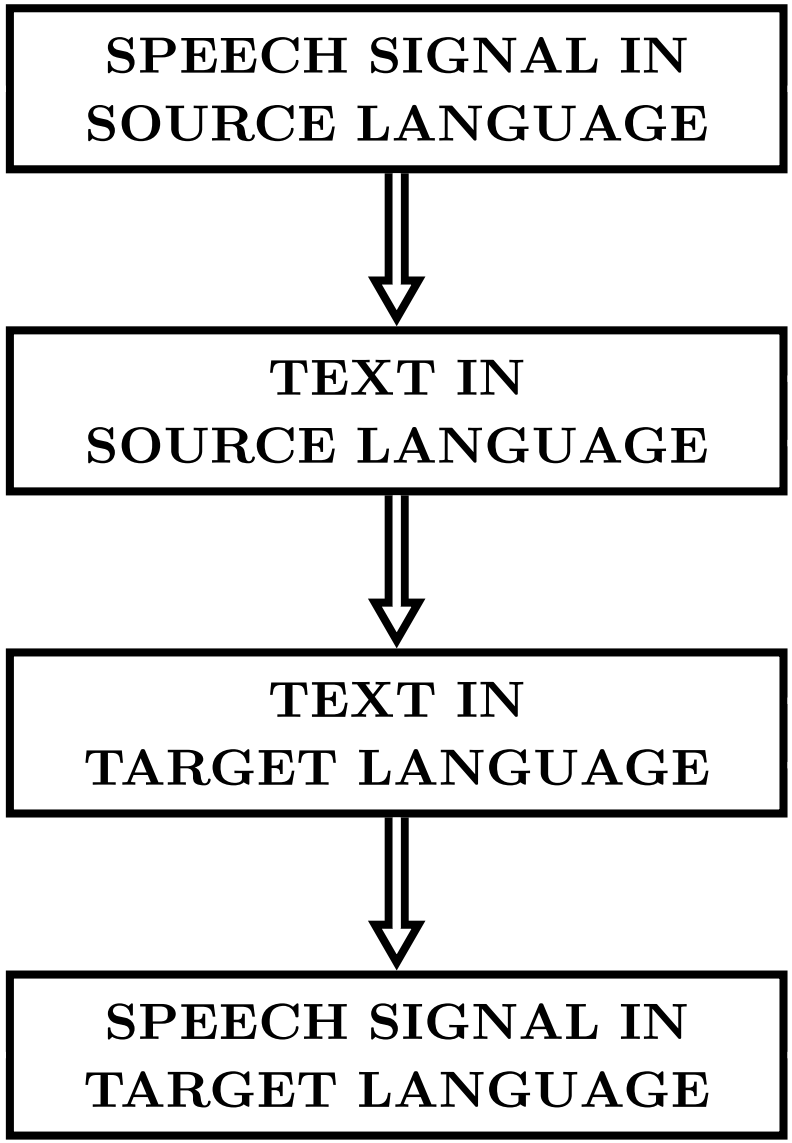
Tasks in Human Language Technology: Speech Translation



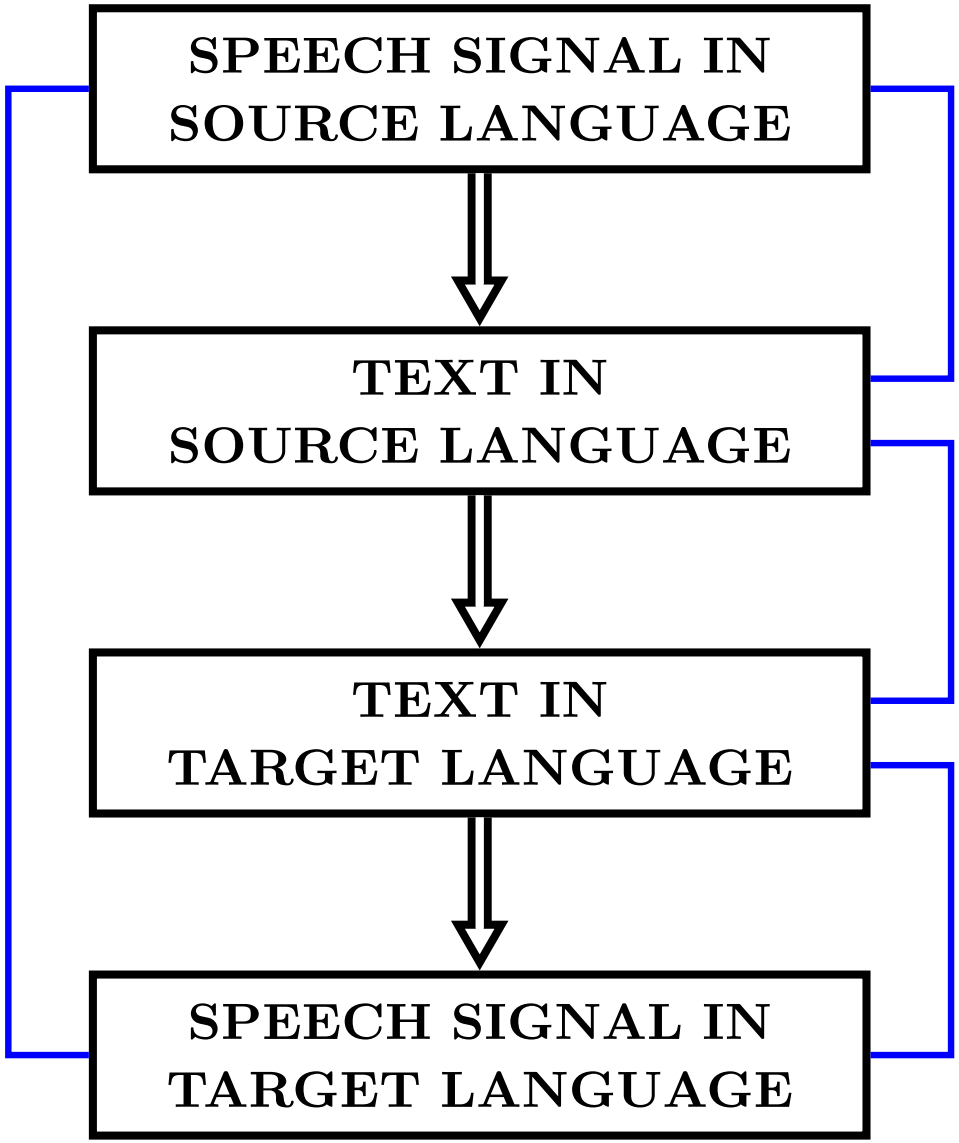
Tasks in Human Language Technology: Speech Translation



Tasks in Human Language Technology: Speech-to-Speech Translation



Tasks in Human Language Technology: Speech-to-Speech Translation



RWTH AACHEN UNIVERSITY

WMT 2017: workshop on machine translation
German – English corpus statistics

WMT 2017		German	English
`train`	Sentences	4.2M	
	Running Words	107M	109M
	Vocabulary	813K	773K
`newtest2015` (dev set)	Sentences	2169	
	Running Words	53K	49K
	Unique words	8022	6706
`newtest2016`	Sentences	2999	
	Running Words	74K	68K
	Unique words	9253	7506
`newtest2017`	Sentences	3004	
	Running Words	73K	68K
	Unique words	9088	7378

HLT

Task WMT 2017: German → English

WMT task: workshop on machine translation
(4M sentence pairs = 100M word pairs)

performance measures:
– BLEU [%]: accuracy measure: "the higher, the better"
– TER [%]: error measure: "the lower, the better"

results for various neural MT approaches:
(RWTH: Wang, Bahar 2017)

	newstest2015 (dev)		newstest2017	
	BLEU	TER	BLEU	TER
Phrase-based system	29.9	54.2	29.8	54.6
Direct HMM: first-order	30.1	51.5	30.4	52.0
zero-order	29.8	51.9	30.6	51.8
Attention: RNN	29.5	52.6	30.1	52.8
RNN 4-enc + 2-dec	31.7	50.7	32.3	50.4
Google's transformer	32.8	49.7	33.5	49.3

Task BOLT 2014: Chinese → English

results [%] for various neural MT approaches:

System	DEV-12		P1R6	
	BLEU	TER	BLEU	TER
Phrase-based system	18.1	68.4	17.3	66.8
Direct HMM: first-order	18.9	67.3	18.8	65.4
zero-order	18.8	67.2	19.0	65.4
Attention: RNN	19.8	65.7	20.0	64.8
RNN 4-enc + 2-dec	20.8	65.4	20.0	65.6
Google's transformer	23.1	62.4	22.5	62.1

Task: WMT 2016 English → Romanian

RWTH AACHEN UNIVERSITY

results [%] for various neural MT approaches:

System	newsdev2016		newstest2016	
	BLEU	TER	BLEU	TER
Phrase-based system	23.9	60.4	24.5	59.3
Direct HMM: first-order	24.9	57.5	24.5	58.0
zero-order	24.2	58.4	24.8	57.8
Attention: RNN	25.1	57.6	26.0	56.7
Google's transformer	27.4	55.5	27.9	54.6

authors: Alkhouli, Bahar, Wang
papers at WMT, EMNLP and ACL

HLT

References

References

[Bahdanau & Cho+ 15] D. Bahdanau, K. Cho, Y. Bengio: Neural machine translation by jointly learning to align and translate. Int. Conf. on Learning and Representation (ICLR), San Diego, CA, May 2015.

[Bahl & Jelinek+ 83] L. R. Bahl, F. Jelinek, R. L. Mercer: A Maximum Likelihood Approach to Continuous Speech Recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 5, pp. 179-190, March 1983.

[Bahl & Brown+ 86] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer: Maximum mutual information estimation of hidden Markov parameters for speech recognition. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Tokyo, pp.49-52, April 1986.

[Beck & Schlüter+ 15] E. Beck, R. Schlüter, H. Ney: Error Bounds for Context Reduction and Feature Omission, Interspeech, Dresden, Germany, Sep. 2015.

[Bengio & Ducharme+ 00] Y. Bengio, R. Ducharme, P. Vincent: A neural probabilistic language model. Advances in Neural Information Processing Systems (NIPS), pp. 933-938, Denver, CO, USA, Nov. 2000.

[Botros & Irie+ 15] R. Botros, K. Irie, M. Sundermeyer, H. Ney: On Efficient Training of Word Classes and Their Application to Recurrent Neural Network Language Models. Interspeech, pp.1443-1447, Dresden, Germany, Sep. 2015.

[Bourlard & Wellekens 89] H. Bourlard, C. J. Wellekens: 'Links between Markov Models and Multilayer Perceptrons', in D.S. Touretzky (ed.): "Advances in Neural Information Processing Systems I", Morgan Kaufmann Pub., San Mateo, CA, pp.502-507, 1989.

[Bridle 89] J. S. Bridle: Probabilistic Interpretation of Feedforward Classification Network Outputs with Relationships to Statistical Pattern Recognition, in F. Fogelman-Soulie, J. Herault (eds.): 'Neuro-computing: Algorithms, Architectures and Applications', NATO ASI Series in Systems and Computer Science, Springer, New York, 1989.

[Brown & Della Pietra+ 93] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, R. L. Mercer: Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, Vol. 19.2, pp. 263-311, June 1993.

[Castano & Vidal+ 93] M.A. Castano, E. Vidal, F. Casacuberta: Inference of stochastic regular languages through simple recurrent networks. IEE Colloquium on Grammatical Inference: Theory, Applications and Alternatives, pp. 16/1-6, Colchester, UK, April 1993.

[Castano & Casacuberta 97] M. Castano, F. Casacuberta: A connectionist approach to machine translation. European Conf. on Speech Communication and Technology (Eurospeech), pp. 91–94, Rhodes, Greece, Sep. 1997.

[Castano & Casacuberta+ 97] M. Castano, F. Casacuberta, E. Vidal: Machine translation using neural networks and finite-state models. Int. Conf. on Theoretical and Methodological Issues in Machine Translation (TMI), pp. 160-167, Santa Fe, NM, USA, July 1997.

[Dahl & Ranzato+ 10] G. E. Dahl, M. Ranzato, A. Mohamed, G. E. Hinton: Phone recognition with the mean-covariance restricted Boltzmann machine. Advances in Neural Information Processing Systems (NIPS) 23, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, Eds. Cambridge, MA, MIT Press, 2010, pp. 469-477.

[Dahl & Yu+ 12] G. E. Dahl, D. Yu, L. Deng, A. Acero: Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Tran. on Audio, Speech and Language Processing, Vol. 20, No. 1, pp. 30-42, Jan. 2012.

[Dehak & Kenny+ 11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet: Front-End Factor Analysis for Speaker Verification IEEE Trans. on audio, speech, and language processing, pp. 788-798, Vol. 19, No. 4, May 2011.

[Devlin & Zbib+ 14] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, J. Makhoul: Fast and Robust Neural Network Joint Models for Statistical Machine Translation. Annual Meeting of the ACL, pp. 1370–1380, Baltimore, MA, June 2014.

[Forcada & Carrasco 05] M. L. Forcada, R. C. Carrasco: Learning the initial state of a second-order recurrent neural network during regular language inference. Neural Computation, Vol. 7, No. 5, pp. 923-930, Sep. 2005.

[Fontaine & Ris+ 97] V. Fontaine, C. Ris, J.-M. Boite: Nonlinear discriminant analysis for improved speech recognition. Eurospeech, Rhodes, Greece, Sep. 1997.

[Fritsch & Finke+ 97] J. Fritsch, M. Finke, A. Waibel: Adaptively Growing Hierarchical Mixtures of Experts. NIPS, Advances in Neural Information Processing Systems 9, MIT Press, pp. 459-465, 1997.

[Gemello & Manai+ 06] R. Gemello, F. Mana, S. Scanzio, P. Lafac, R. De Mori: Adaptation of Hybrid ANN/HMM Models Using Linear Hidden Transformations and Conservative Training. IEEE Int. Conf. on Acoustics Speech and Signal Processing Proceedings, Toulouse, 2006.

[Gers & Schmidhuber+ 00] F. A. Gers, J. Schmidhuber, F. Cummin: Learning to forget: Continual prediction with LSTM. Neural computation, Vol 12, No. 10, pp. 2451-2471, 2000.

[Gers & Schraudolph+ 02] F. A. Gers, N. N. Schraudolph, J. Schmidhuber: Learning precise timing with LSTM recurrent networks. Journal of Machine Learning Research, Vol. 3, pp. 115-143, 2002.

[Graves & Fernandez+ 06] A. Graves, S. Fernandez, F Gomez, J. Schmidhuber: Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. Int. Conf. on Machine Learning, Pittsburgh, PA, pp. 369-376, 2006.

[Graves & Schmidhuber 09] A. Graves, J. Schmidhuber: Offline handwriting recognition with multidimensional recurrent neural networks. NIPS 2009.

[Grezl & Fousek 08] F. Grezl, P. Fousek: Optimizing bottle-neck features for LVCSR. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 4729-4732, Las Vegas, NV, March 2008.

[Haffner 93] P. Haffner: Connectionist Speech Recognition with a Global MMI Algorithm. 3rd Europ. Conf. on Speech Communication and Technology (Eurospeech'93), Berlin, Germany, Sep. 1993.

[Heigold & Macherey05] G. Heigold, W. Macherey, R. Schlüter, H. Ney: Minimum Exact Word Error Training. IEEE ASRU workshop, pp. 186-190, San Juan, Puerto Rico, Nov. 2005.

[Heigold & Schlüter12] G. Heigold, R. Schlüter, H. Ney, S. Wiesler: Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance. IEEE Signal Processing Magazine, vol. 29, no. 6, pp. A58-69, Nov. 2012.

[Hermansky & Ellis+ 00] H. Hermansky, D. W. Ellis, S. Sharma: Tandem connectionist feature extraction for conventional HMM systems. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 1635-1638, Istanbul, Turkey, June 2000.

[Hinton & Osindero+ 06] G. E. Hinton, S. Osindero, Y. Teh: A fast learning algorithm for deep belief nets. Neural Computation, Vol. 18, No. 7, pp. 1527-1554, July 2006.

[Hochreiter & Schmidhuber 97] S. Hochreiter, J. Schmidhuber: Long short-term memory. Neural Computation, Vol. 9, No. 8, pp. 1735–1780, Nov. 1997.

[Ivakhnenko 71] A. G. .Ivakhnenko: Polynomial theory of complex systems. IEEE Transactions on Systems, Man and Cybernetics, Vol. 1, No. 4, pp. 364-378, Oct. 1971.

[Klakow & Peters 02] D. Klakow, J. Peters: Testing the correlation of word error rate and perplexity. Speech Communication, pp. 19–28, 2002.

[Koehn & Och+ 03] P. Koehn, F. J. Och, D. Marcu: Statistical Phrase-Based Translation. HLT-NAACL 2003, pp. 48-54, Edmonton, Canada, May-June 2003.

[Le & Allauzen+ 12] H.S. Le, A. Allauzen, F. Yvon: Continuous space translation models with neural networks. NAACL-HLT 2012, pp. 39-48, Montreal, QC, Canada, June 2002.

[LeCun & Bengio+ 94] Y. LeCun, Y. Bengio: Word-level training of a handwritten word recognizer based on convolutional neural networks. Int. Conf. on Pattern Recognition, Jerusalem, Israel, pp. 88-92, Oct. 1994.

[Makhoul & Schwartz 94] J. Makhoul, R Schwartz: State of the Art in Continuous Speech Recognition. Chapter 14, pp. 165-198, in D. B. Roe, J. G. Wilpon (Editors): Voice Communication Between Humans and Machines. National Academy of Sciences, 1994.

[Miao & Metze 15] Y. Miao. F Metze: On speaker adaptation of long short-term memory recurrent neural networks. Interspeech, Dresden, Germany, 2015.

[Mikolov & Karafiat+ 10] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, S. Khudanpur: Recurrent neural network based language model. Interspeech, pp. 1045-1048, Makuhari, Chiba, Japan, Sep. 2010.

[Mohamed & Dahl+ 09] A. Mohamed, G. Dahl, G. Hinton: Deep belief networks for phone recognition. NIPS Workshop Deep Learning for Speech Recognition and Related Applications, 2009.

[Nakamura & Shikano 89] M. Nakamura, K. Shikano: A Study of English Word Category Prediction Based on Neural Networks. ICASSP 89, p. 731-734, Glasgow, UK, May 1989.

[Neco & Forcada 97] R. P. Neco, M. L. Forcada: Asynchronous translations with recurrent neural nets. IEEE Int. Conf. on Neural Networks, pp. 2535-2540, June 1997.

[Normandin & Cardin+ 94] Y. Normandin, R. Cardin, R. De Mori: High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation. IEEE Trans. on Speech and Audio Processing, vol. 2, no. 2, pp. 299-311, April 1994.

[Ney 03] H. Ney: On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition. First Iberian Conf. on Pattern Recognition and Image Analysis, Puerto de Andratx, Spain, Springer LNCS Vol. 2652, pp. 636-645, June 2003.

[Och & Ney 03] F. J. Och, H. Ney: A Systematic Comparison of Various Alignment Models. *Computational Linguistics,* Vol. 29, No. 1, pp. 19-51, March 2003.

[Och & Ney 04] F. J. Och, H. Ney: The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, Vol. 30, No. 4, pp. 417-449, Dec. 2004.

[Och & Tillmann+ 99] F. J. Och, C. Tillmann, H. Ney: Improved Alignment Models for Statistical Machine Translation. Joint ACL/SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora, College Park, MD, pp. 20-28, June 1999.

[Patterson & Womack 66] J. D. Patterson, B. F. Womack: An Adaptive Pattern Classification Scheme. IEEE Trans. on Systems, Science and Cybernetics, Vol.SSC-2, pp.62-67, Aug. 1966.

[Povey & Woodland 02] D. Povey, P.C. Woodland: Minimum phone error and I-smoothing for improved discriminative training. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 105–108, Orlando, FL, May 2002.

[Printz & Olsen 02] H. Printz, P. A. Olsen: Theory and practice of acoustic confusability. Computer Speech and Language, pp. 131–164, Jan. 2002.

[Robinson 94] A. J. Robinson: An Application of Recurrent Nets to Phone Probability Estimation. IEEE Trans. on Neural Networks, Vol. 5, No. 2, pp. 298-305, March 1994.

[Schlüter & Nussbaum+ 11] R. Schlüter, M. Nussbaum-Thom, H. Ney: On the Relationship between Bayes Risk and Word Error Rate in ASR. IEEE Trans. on Audio, Speech, and Language Processing, vol. 19, no. 5, p. 1103-1112, July 2011.

[Schlüter & Nussbaum+ 12] R. Schlüter, M. Nussbaum-Thom, H. Ney: Does the Cost Function Matter in Bayes Decision Rule? IEEE Trans. PAMI, No. 2, pp. 292–301, Feb. 2012.

[Schlüter & Nussbaum-Thom+ 13] R. Schlüter, M. Nußbaum-Thom, E. Beck, T. Alkhouli, H. Ney: Novel Tight Classification Error Bounds under Mismatch Conditions based on f-Divergence. IEEE Information Theory Workshop, pp. 432–436, Sevilla, Spain, Sep. 2013.

[Schlüter & Scharrenbach+ 05] R. Schlüter, T. Scharrenbach, V. Steinbiss, H. Ney: Bayes Risk Minimization using Metric Loss Functions Interspeech, pages 1449-1452, Lisboa, Portugal, Sep. 2005.

[Schuster & Paliwal 97] M. Schuster, K. K. Paliwal: Bidirectional Recurrent Neural Networks. IEEE Trans. on Signal Processing, Vol. 45, No. 11, pp. 2673-2681, Nov. 1997.

[Schwenk 07] H. Schwenk: Continuous space language models. Computer Speech and Language, Vol. 21, No. 3, pp. 492–518, July 2007.

[Schwenk 12] H. Schwenk: Continuous Space Translation Models for Phrase-Based Statistical Machine Translation. 24th Int. Conf. on Computational Linguistics (COLING), Mumbai, India, pp. 1071–1080, Dec. 2012.

[Schwenk & Costa-jussa+ 07] H. Schwenk , M. R. Costa-jussa, J. A. R. Fonollosa: Smooth bilingual n-gram translation. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 430–438, Prague, June 2007.

[Schwenk & Déchelotte+ 06] H. Schwenk, D. Déchelotte, J. L. Gauvain: Continuous Space Language Models for Statistical Machine Translation. COLING/ACL 2006, pp. 723–730, Sydney, Australia July 2006.

[Seide & Li+ 11] F. Seide, G. Li, D. Yu: Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. Interspeech, pp. 437-440, Florence, Italy, Aug. 2011.

[Solla & Levin+ 88] S. A. Solla, E. Levin, M. Fleisher: Accelerated Learning in Layered Neural Networks. Complex Systems, Vol.2, pp. 625-639, 1988.

[Stolcke & Grezl+ 06] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, D. Vergyri: Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Toulouse, France, May 2006.

[Sundermeyer & Alkhouli+ 14] M. Sundermeyer, T. Alkhouli, J. Wuebker, H. Ney: Translation Modeling with Bidirectional Recurrent Neural Networks. Conf. on Empirical Methods in Natural Language Processing (EMNLP), pp. 14–25, Doha, Qatar, Oct. 2014.

[Sundermeyer & Ney+ 15] M. Sundermeyer, H. Ney, R. Schlüter: From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Trans. on Audio, Speech, and Language Processing, Vol. 23, No. 3, pp. 13–25, March 2015.

[Sundermeyer & Schlüter+ 12] M. Sundermeyer, R. Schlüter, H. Ney: LSTM neural networks for language modeling. Interspeech, pp. 194–197, Portland, OR, USA, Sep. 2012.

[Tüske & Plahl+ 11] Z. Tüske, C. Plahl, R. Schlüter: A study on speaker normalized MLP features in LVCSR. Interspeech, pp. 1089-1092, Florence, Italy, Aug. 2011.

[Utgoff & Stracuzzi 02] P. E. Utgoff, D. J. Stracuzzi: Many-layered learning. Neural Computation, Vol. 14, No. 10, pp. 2497-2539, Oct. 2002.

[Valente & Vepa+ 07] F. Valente, J. Vepa, C. Plahl, C. Gollan, H. Hermansky, R. Schlüter: Hierarchical Neural Networks Feature Extraction for LVCSR system. Interspeech, pp. 42-45, Antwerp, Belgium, Aug. 2007.

[Vapnik 98] Vapnik: Statistical Learning Theory. Addison-Wesley, 1998.

[Vaswani & Zhao+ 13] A. Vaswani, Y. Zhao, V. Fossum, D. Chiang: Decoding with Large-Scale Neural Language Models Improves Translation. Conf. on Empirical Methods in Natural Language Processing (EMNLP, pp. 1387–1392, Seattle, Washington, USA, Oct. 2013.

[Vogel & Ney+ 96] S. Vogel, H. Ney, C. Tillmann: HMM-based word alignment in statistical translation. Int. Conf. on Computational Linguistics (COLING), pp. 836-841, Copenhagen, Denmark, Aug. 1996.

[Waibel & Hanazawa+ 88] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. L. Lang: Phoneme Recognition: Neural Networks vs. Hidden Markov Models. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), New York, NY, pp.107-110, April 1988.

[Xu & Povey+ 10] H. Xu, D. Povey, L. Mangu, J. Zhu: Minimum Bayes Risk Decoding and System Combination Based on a Recursion for Edit Distance. Computer Speech and Language, Sep. 2010.

[Zens & Och+ 02] R. Zens, F. J. Och, H. Ney: Phrase-Based Statistical Machine Translation. 25th Annual German Conf. on AI, pp. 18–32, LNAI, Springer 2002.

THE END
RAAI Moscow
06-Jul-2019