Neural MT

Announcements

- HW2 directory structure penalty to be removed due to grading inconsistencies.
  - Those who lost 15 points will gain 15 points

- Dan Jurafsky will attend the beginning of class next Tuesday
  - Be prepared with questions. Your chance!!!

- Rupal Patel: Monday, Dec. 4^{th}, 11:30, Davis

- **Data Science Institute Colloquium Series Event: DAN JURAFSKY, STANFORD UNIVERSITY | Tuesday, December 5th at 5PM in Davis Auditorium (412 CEPSR)**

- **"Does This Vehicle Belong to You?" Processing the Language of Policing for Improving Police-Community Relations**
-
- ABSTRACT
- Police body-worn cameras have the potential to play an important role in understanding and improving police-community relations. In this talk I describe a series of studies conducted by our large interdisciplinary team at Stanford that use speech and natural language processing on body-camera recordings to model the interactions between police officers and community members in traffic stops. We use text and speech features to automatically measure linguistic aspects of the interaction, from discourse factors like conversational structure to social factors like respect. I describe the differences we find in the language directed toward black versus white community members, and offer suggestions for how these findings can be used to help improve the fraught relations between police officers and the communities they serve.

Today

- Multilingual Challenges for MT
- MT Approaches
  - Statistical
  - Neural net (Thursday)
- MT Evaluation

MT Evaluation

- More art than science
- Wide range of Metrics/Techniques
  - interface, …, scalability, …, faithfulness, ... space/time complexity, … etc.
- Automatic vs. Human-based
  - *Dumb Machines vs. Slow Humans*

Slide from Nizar Habash

Human-based Evaluation Example
Accuracy Criteria

5	contents of original sentence conveyed (might need minor corrections)
4	contents of original sentence conveyed BUT errors in word order
3	contents of original sentence generally conveyed BUT errors in relationship between phrases, tense, singular/plural, etc.
2	contents of original sentence not adequately conveyed, portions of original sentence incorrectly translated, missing modifiers
1	contents of original sentence not conveyed, missing verbs, subjects, objects, phrases or clauses

Slide from Nizar Habash

Human-based Evaluation Example
Fluency Criteria

5	clear meaning, good grammar, terminology and sentence structure
4	clear meaning BUT bad grammar, bad terminology or bad sentence structure
3	meaning graspable BUT ambiguities due to bad grammar, bad terminology or bad sentence structure
2	meaning unclear BUT inferable
1	meaning absolutely unclear

Slide from Nizar Habash

Today: Crowdsourcing

- Amazon Mechanical Turk or CrowdFlower
- Create a HIT for each sentence
- Get multiple workers to rate
- Pay .01 to .10 per hit
- Complete an evaluation in hours (vs days/ weeks)
- *Ethics?*

Automatic Evaluation Example
Bleu Metric
(Papineni et al 2001)

- Bleu
  - *BiLingual Evaluation Understudy*
  - Modified n-gram precision with length penalty
  - Quick, inexpensive and language independent
  - Correlates highly with human evaluation
  - Bias against synonyms and inflectional variations

Slide from Nizar Habash

Automatic Evaluation Example
Bleu Metric

Test Sentence

colorless green ideas sleep furiously

Gold Standard References

all dull jade ideas sleep irately
drab emerald concepts sleep furiously
colorless immature thoughts nap angrily

Slide from Nizar Habash

Automatic Evaluation Example
Bleu Metric

Test Sentence

colorless green ideas sleep furiously

Gold Standard References

all dull jade ideas sleep irately
drab emerald concepts sleep furiously
colorless immature thoughts nap angrily

Unigram precision = 4/5

Slide from Nizar Habash

Automatic Evaluation Example
Bleu Metric

Test Sentence

colorless green
green ideas
ideas sleep
sleep furiously

Gold Standard References

all dull jade ideas sleep irately
drab emerald concepts sleep furiously
colorless immature thoughts nap angrily

Unigram precision = 4 / 5 = 0.8
Bigram precision = 2 / 4 = 0.5

Bleu Score = $(a_1 \, a_2 \ldots a_n)^{1/n}$
$= (0.8 \times 0.5)^{½} = 0.6325 \rightarrow 63.25$

Slide from Nizar Habash

BLEU scores for 110 translation systems trained on Europarl

Source Language	Target Language										
	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	**31.2**	**32.1**	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	**30.1**	**31.1**	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	**30.5**	-	**40.2**	12.5	**32.3**	21.4	**35.9**	23.9
fr	23.7	18.5	26.1	**30.0**	**38.4**	-	12.6	**32.4**	21.1	**35.3**	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	**34.0**	**36.0**	11.0	-	20.0	**31.2**	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	**30.1**	**37.9**	**39.0**	11.9	**32.0**	20.2	-	21.9
sv	**30.3**	18.9	22.8	**30.2**	28.6	29.7	15.3	23.9	21.9	25.9	-

Table 2: BLEU scores for the 110 translation systems trained on the Europarl corpus

Koehn, MT Summit, 2005
http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf

Language	From	Into	Diff
Danish (da)	23.4	23.3	0.0
German (de)	**22.2**	**17.7**	**-4.5**
Greek (el)	23.8	22.9	-0.9
English (en)	**23.8**	**27.4**	**+3.6**
Spanish (es)	26.7	29.6	+2.9
French (fr)	26.1	31.1	+5.1
Finnish (fi)	19.1	12.4	-6.7
Italian (it)	24.3	25.4	+1.1
Dutch (nl)	19.7	20.7	+1.1
Portuguese (pt)	26.1	27.0	+0.9
Swedish (sv)	24.8	22.1	-2.6

Table 3: Average translation scores for systems when translating *from* and *into* a language. Note that German (de) and English (en) are similarly difficult to translate *from*, but English is much easier to translate *into*.

Automatic Evaluation Example
METEOR
(Lavie and Agrawal 2007)

- Metric for Evaluation of Translation with Explicit word Ordering
- Extended Matching between translation and reference
  - Porter stems, wordNet synsets
- Unigram Precision, Recall, parameterized F-measure
- Reordering Penalty
- Parameters can be tuned to optimize correlation with human judgments
- Not biased against "non-statistical" MT systems

Slide from Nizar Habash

Metrics MATR Workshop

- Workshop in AMTA conference 2008
  - Association for Machine Translation in the Americas
- Evaluating evaluation metrics
- Compared 39 metrics
  - 7 baselines and 32 new metrics
  - Various measures of correlation with human judgment
  - Different conditions: text genre, source language, number of references, etc.

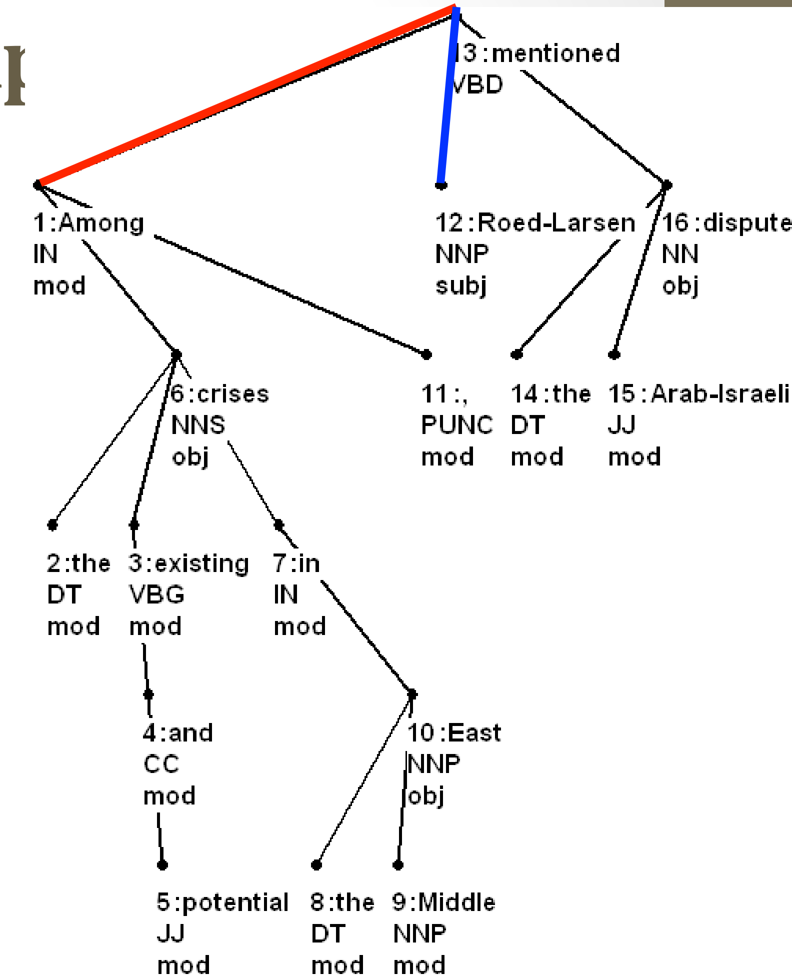
Slide from Nizar Habash

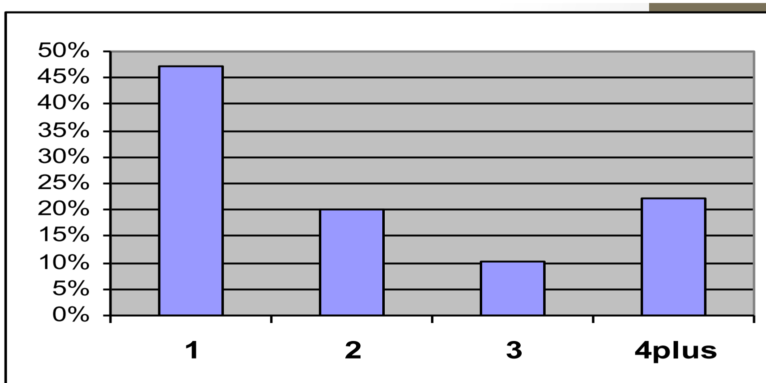
Automatic Evaluation Examp

SEPIA

(Habash and ElKholy 2008)

- A syntactically-aware evaluation metric
  - (Liu and Gildea, 2005; Owczarzak et al., 2007; Giménez and Màrquez, 2007)
- Uses dependency representation
  - MICA parser (Nasr & Rambow 2006)
  - 77% of all structural bigrams are surface n-grams of size 2,3,4
- Includes dependency surface span as a factor in score
  - long-distance dependencies should receive a greater weight than short distance dependencies
    - Higher degree of grammaticality?





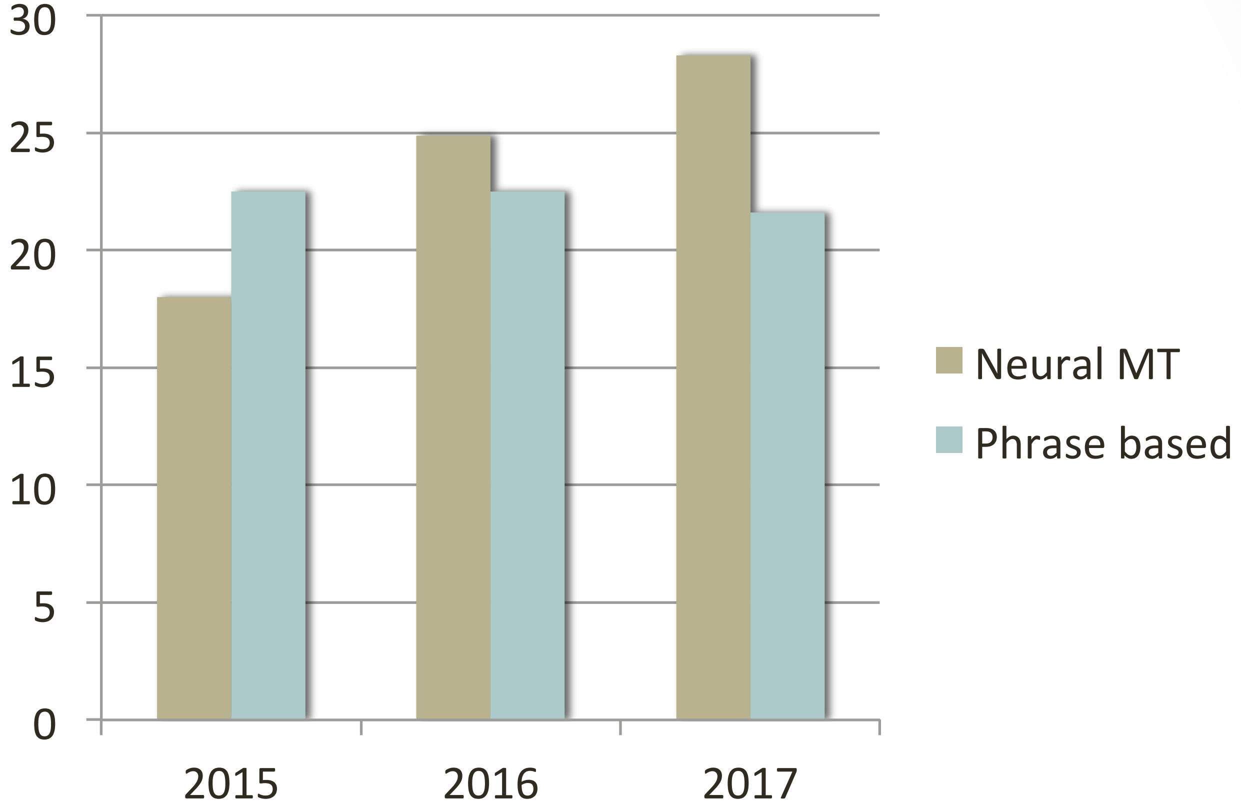
Why do people continue to use BLEU

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

Neural MT takes over

- WMT (Workshop on Machine Translation)
- 2015 – first neural MT, lower bleu results
- 2016: neural MT beats phrase-based and syntax-based



Results from WMT (Workshop on Machine Translation)
German to English
2015: Montreal
2016 and 2017: Edinburgh

WMT 2017

- Tasks
  - News translation
  - Quality estimation
  - Automatic post-editing
  - Metrics
  - Multimodal MT and multilingual image description
  - Biomedical translation

What is being tested in the biomedical task?

Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

News Translation Task

- 7 languages, 14 tasks (from and into English)
  - Chinese
  - Czech
  - German
  - Finnish
  - Latvian
  - Russian
  - Turkish
- Test data: 3000 sentences per language pair except Latvian: 2000 sentences

Training Data

- Europarl
- Common Crawl
- Yandex Russian-English data
- Wikipedia Headlines
- United Nations
- News Commentary V12
- EU Press Release parallel corpus for German, Finnish and Latvian

Submitted Systems

- 103 systems from 31 institutions (no companies)
- Company releases of Neural MT
  - Microsoft: February 2016
  - Systran: August 2016
  - Google: September 2016

Human Evaluation

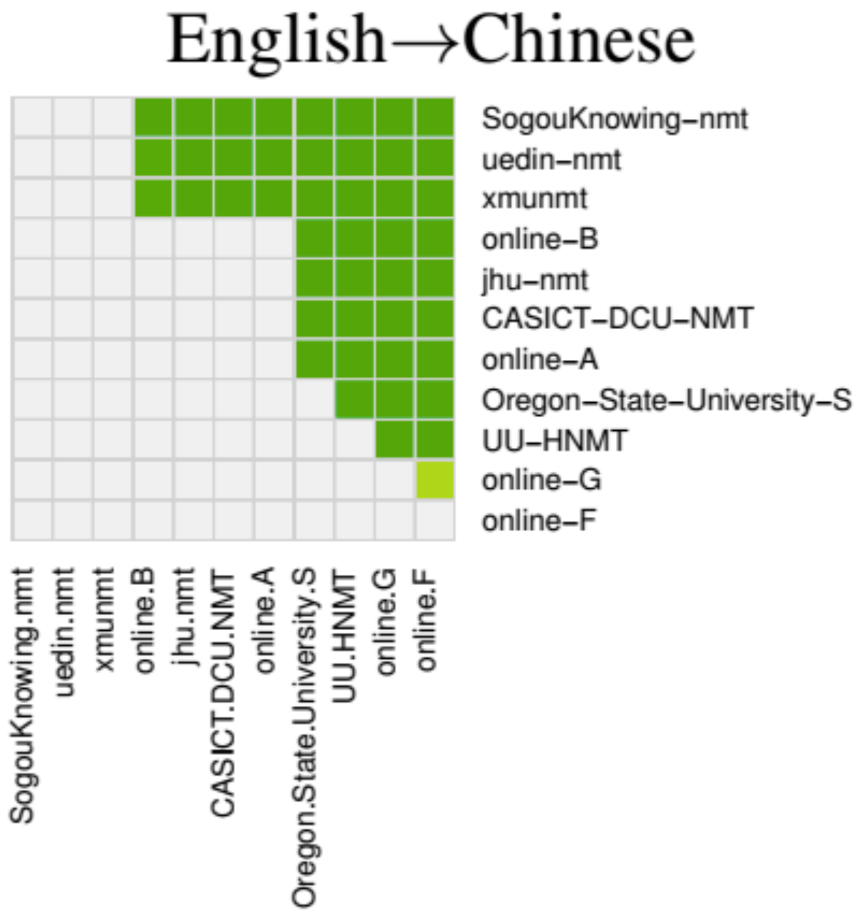
- Assess on adequacy along a 100 point scale (Direct Assessment) (vs Relative Ranking)
  - How adequately does the translation express the meaning of the reference translation?
  - One translation per screen/hit
- 151 individual Researchers
  - 29 different groups
  - Contributed 12,693 translation scores
  - 24 days, 22 hours
- 754 AMT workers
  - Contributed 237,200 scores
  - 47 days, 23 hours

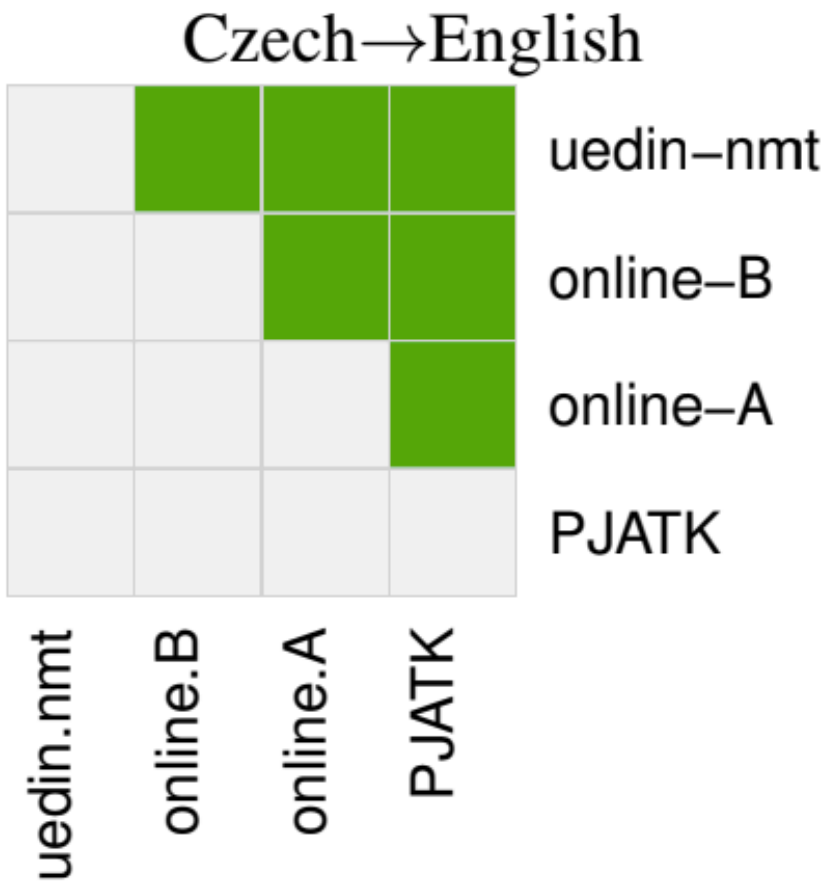
ld we control quality on an Amazon Mechani evaluation? Is it reliable?

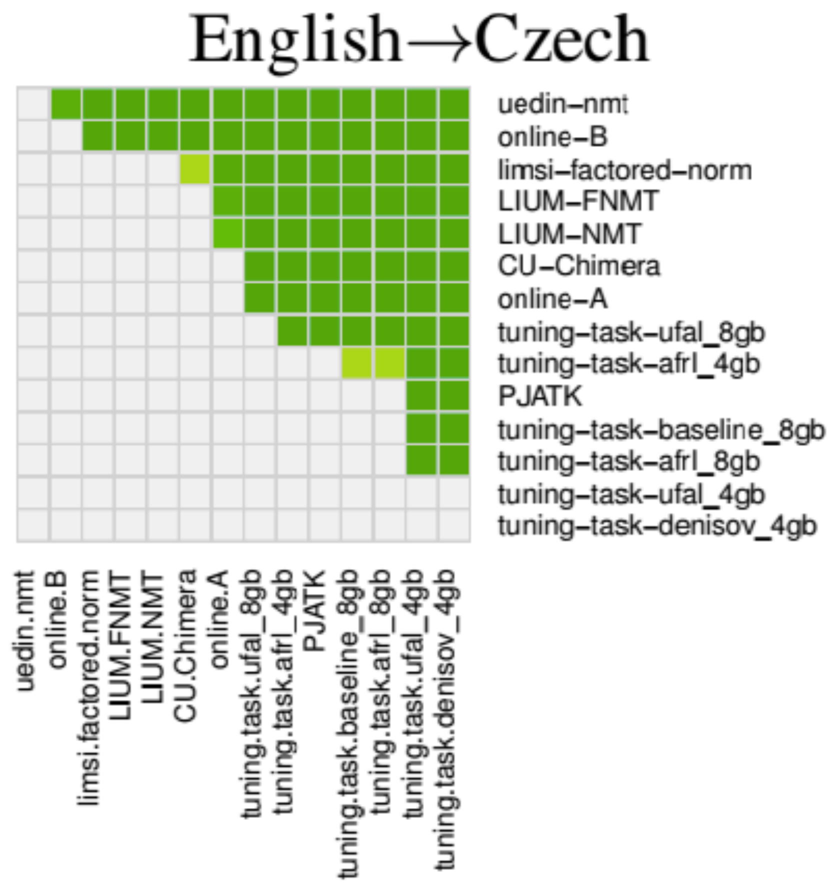
Start the presentation to activate live content

If you see this message in presentation mode, install the add-in or get help at PollEv.com/app

Some Results




Czech→English
uedin–nmt
online–B
online–A
PJATK
uedin.nmt
online.B
online.A
PJATK


English→Czech
uedin–nmt
online–B
limsi–factored–norm
LIUM–FNMT
LIUM–NMT
CU–Chimera
online–A
tuning–task–ufal_8gb
tuning–task–afrl_4gb
PJATK
tuning–task–baseline_8gb
tuning–task–afrl_8gb
tuning–task–ufal_4gb
tuning–task–denisov_4gb
uedin.nmt
online.B
limsi.factored.norm
LIUM.FNMT
LIUM.NMT
CU.Chimera
online.A
tuning.task.ufal_8gb
tuning.task.afrl_4gb
PJATK
tuning.task.baseline_8gb
tuning.task.afrl_8gb
tuning.task.ufal_4gb
tuning.task.denisov_4gb

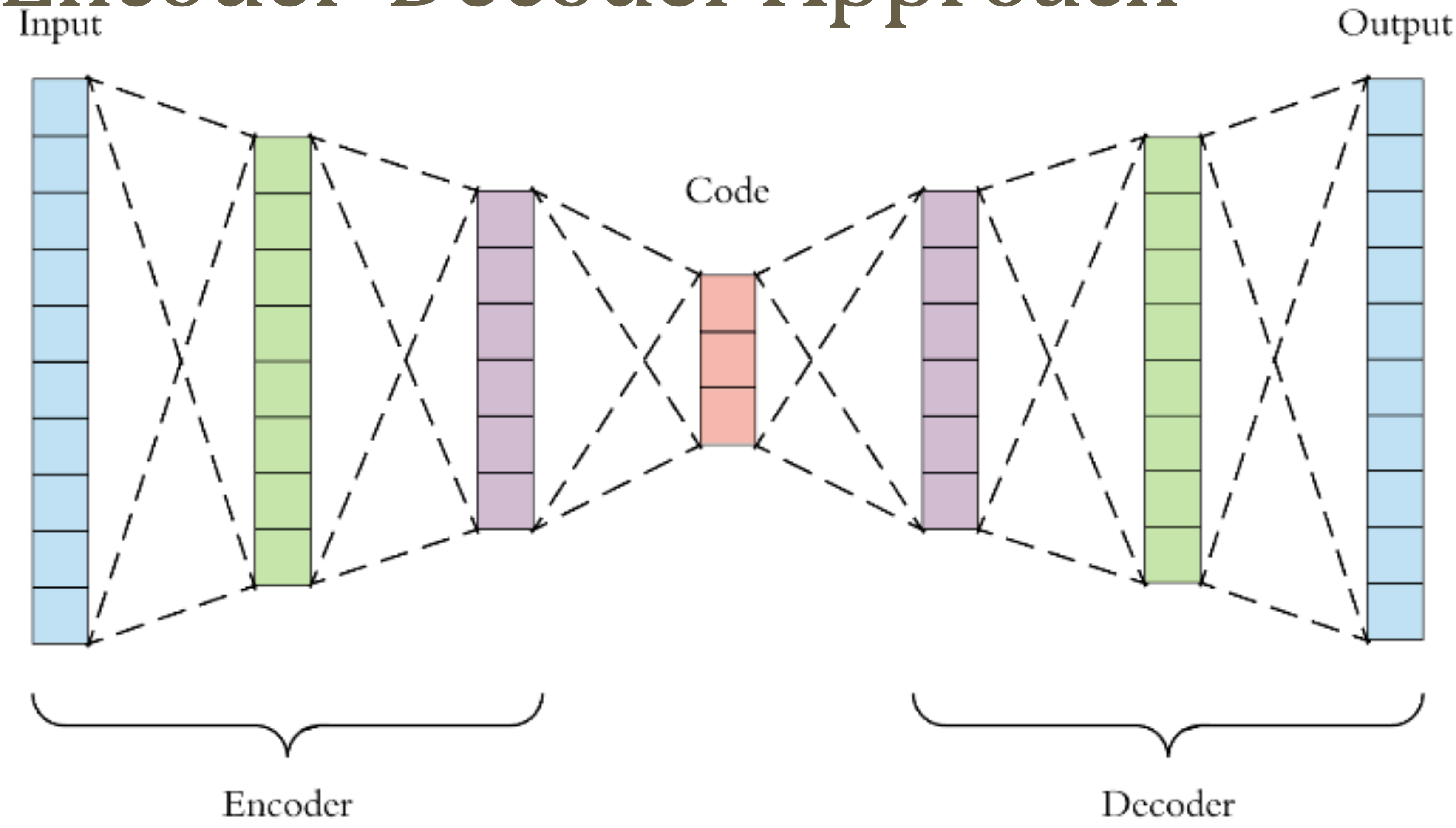
Today

- Multilingual Challenges for MT

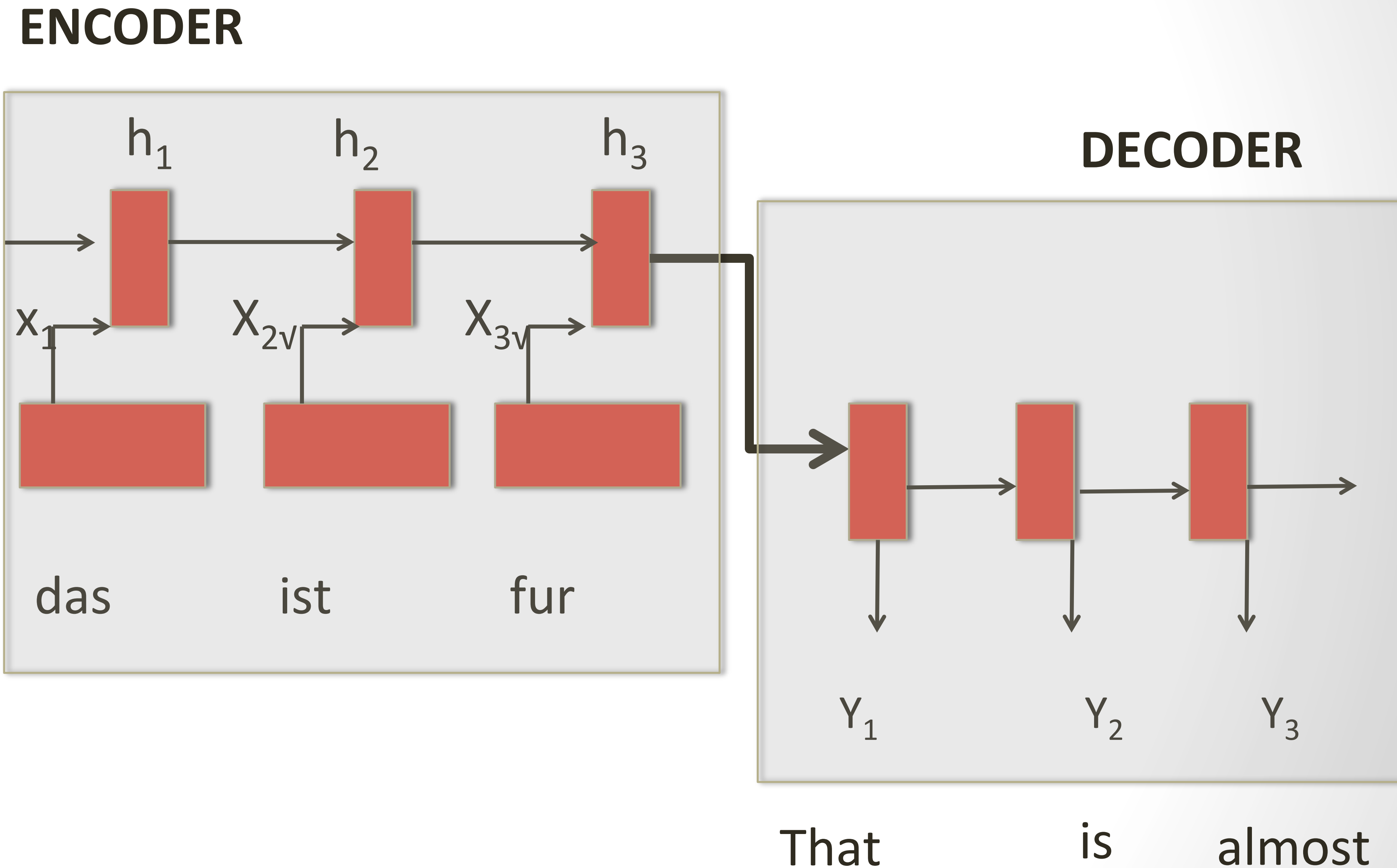
- MT Approaches
  - Statistical
  - Neural net (Thursday)

- MT Evaluation

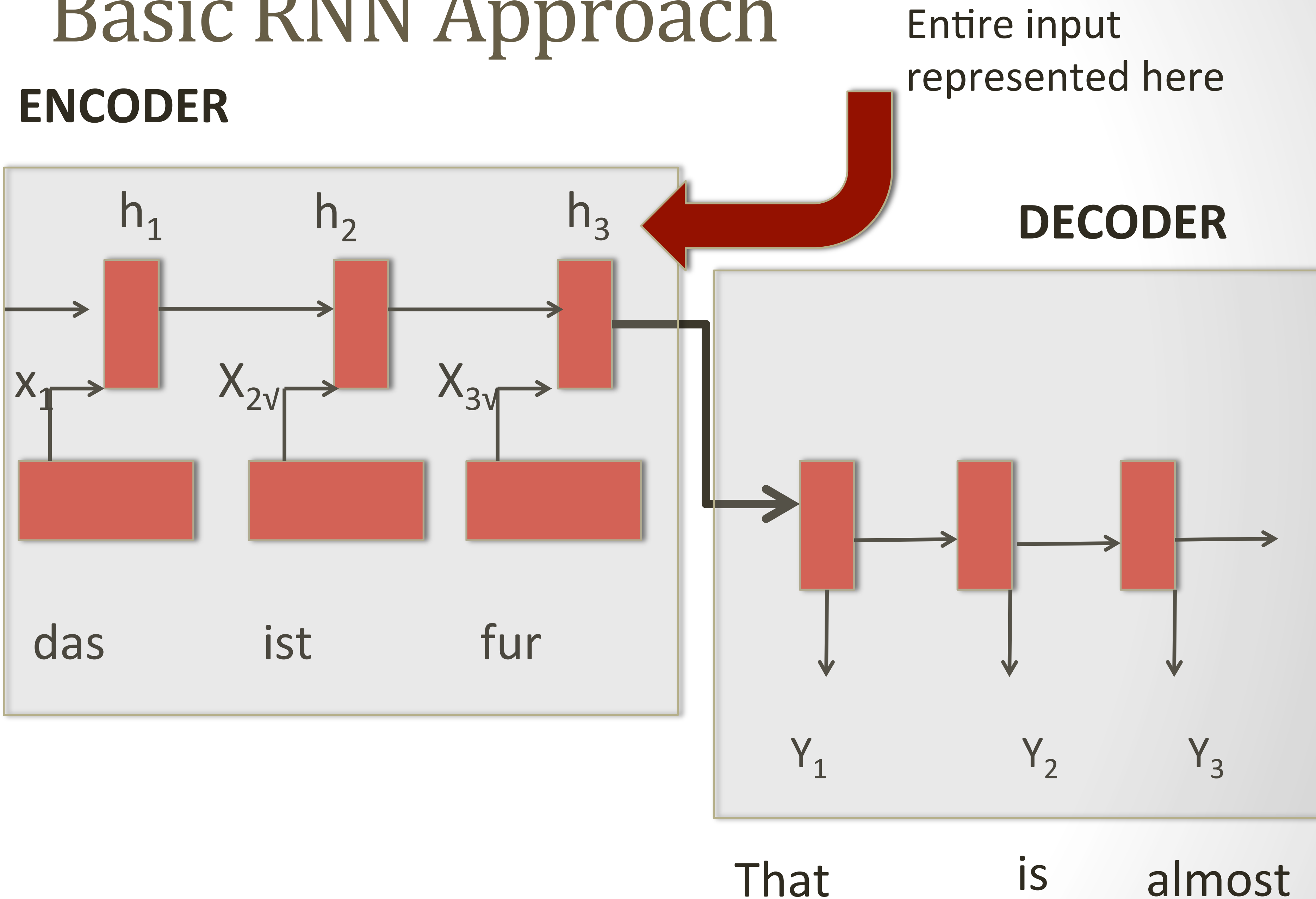
Encoder-Decoder Approach



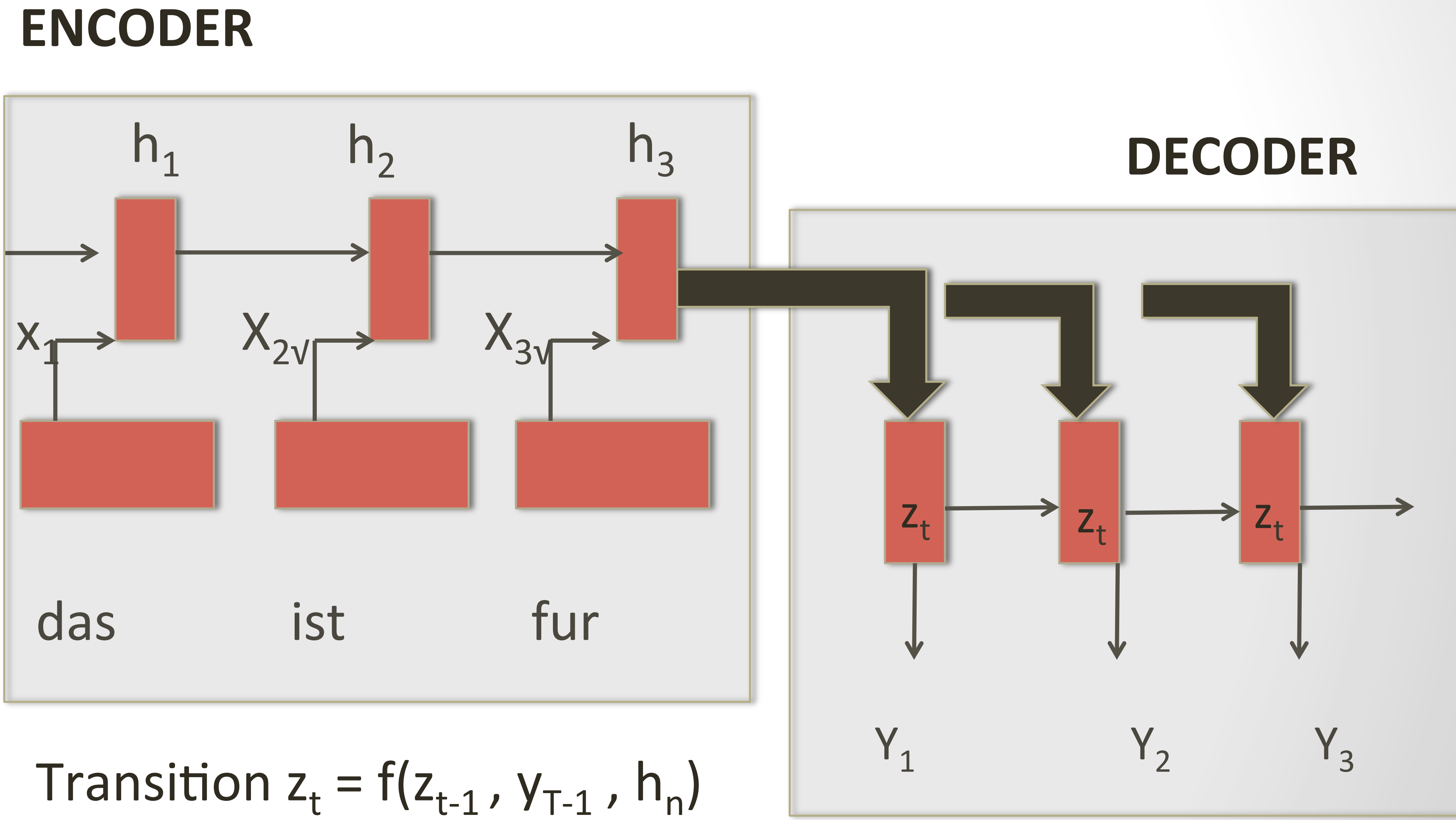
Basic RNN Approach



Basic RNN Approach



Recurrent decoder *but*



Transition $z_t = f(z_{t-1}, y_{T-1}, h_n)$
Backpropagation $= \Sigma_t \, \delta z_t/\delta h$

That is almost

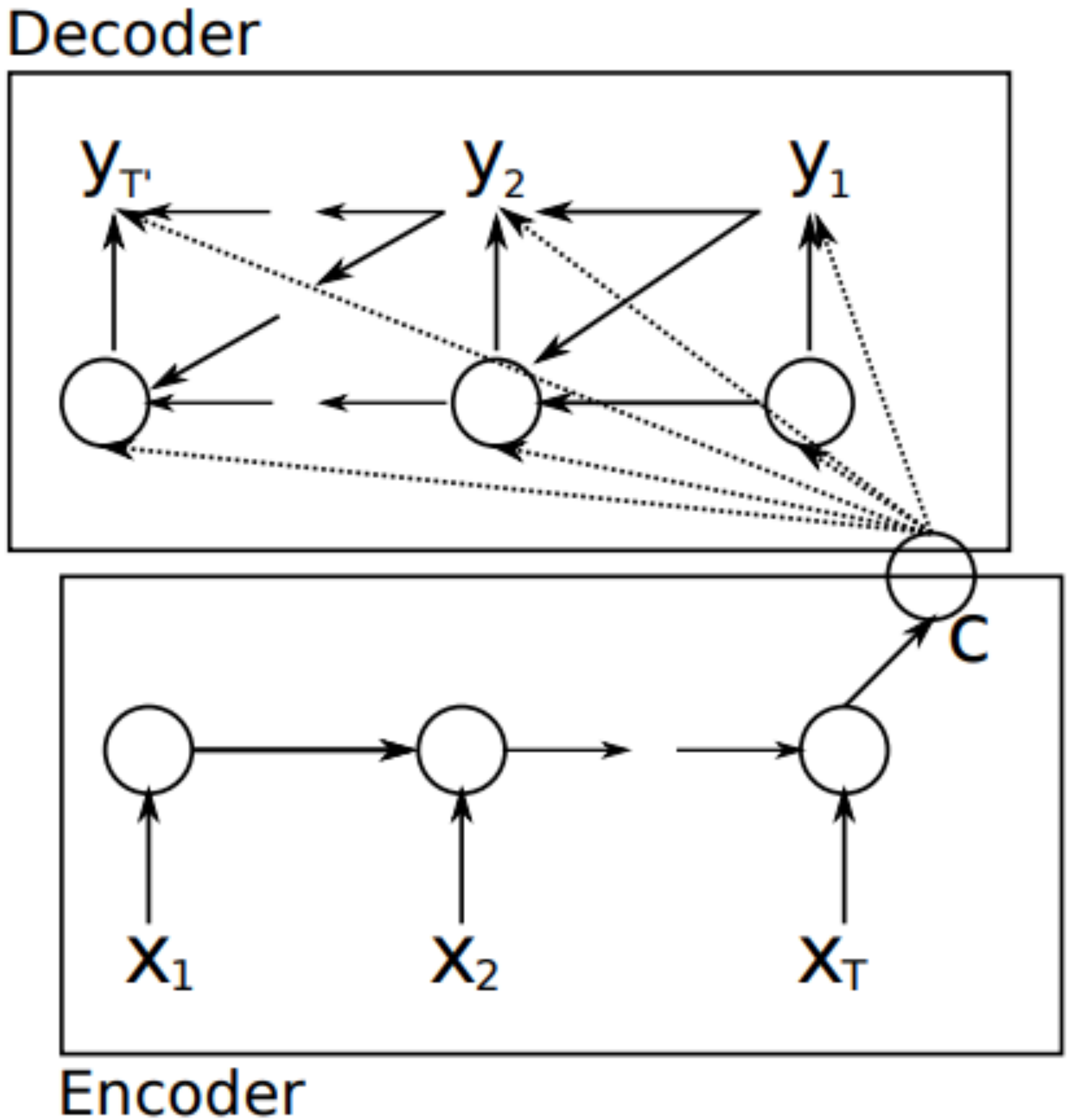


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

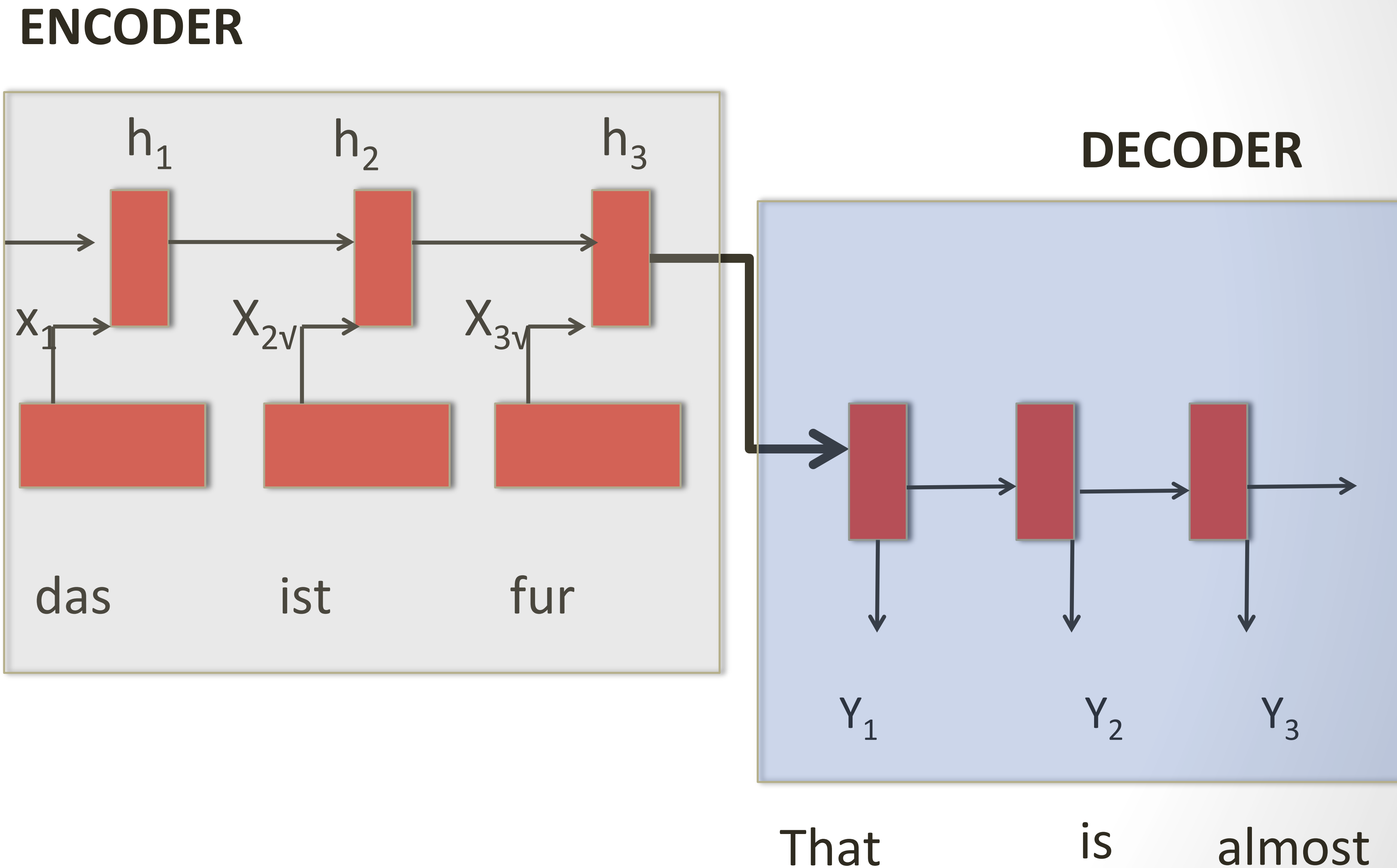
Cho et al 2014

Results for Long Frequent Phrases

		RNN Encoder–Decoder
at the end of the	[a la fin de la] [ŕ la fin des années] [être supprimés à la fin de la]	[à la fin du] [à la fin des] [à la fin de la]
for the first time	[**r** © pour la premi**r**ëre fois] [été donnés pour la première fois] [été commémorée pour la première fois]	[pour la première fois] [pour la première fois ,] [pour la première fois que]
in the United States and	[? aux ?tats-Unis et] [été ouvertes aux États-Unis et] [été constatées aux États-Unis et]	[aux Etats-Unis et] [des Etats-Unis et] [des États-Unis et]
, as well as	[?s , qu’] [?s , ainsi que] [?re aussi bien que]	[, ainsi qu’] [, ainsi que] [, ainsi que les]
one of the most	[?t ?l’ un des plus] [?l’ un des plus] [être retenue comme un de ses plus]	[l’ un des] [le] [un des]

Cho et al 2014

Other Variants: Train weights separately



Also Useful

- Train stacked RNNS using multiple layers
- Use a bidirectional encoder
  - This can help in remembering the early part of the source input sentence
- Train the input sequence in reverse order: $S_1S_2S_3$ -> $T_1T_2T_3$ would be trained as $S_3S_2S_1$ -> $T_1T_2T_3$
  - Why?

Replacing RNN with LSTM improves performance further

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Table 1: The performance of the LSTM on WMT'14 English to French test set (ntst14). Note that an ensemble of 5 LSTMs with a beam of size 2 is cheaper than of a single LSTM with a beam of size 12.

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
State of the art [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	$\sim$45

Table 2: Methods that use neural networks together with an SMT system on the WMT'14 English to French test set (ntst14).

Aligning and Translating

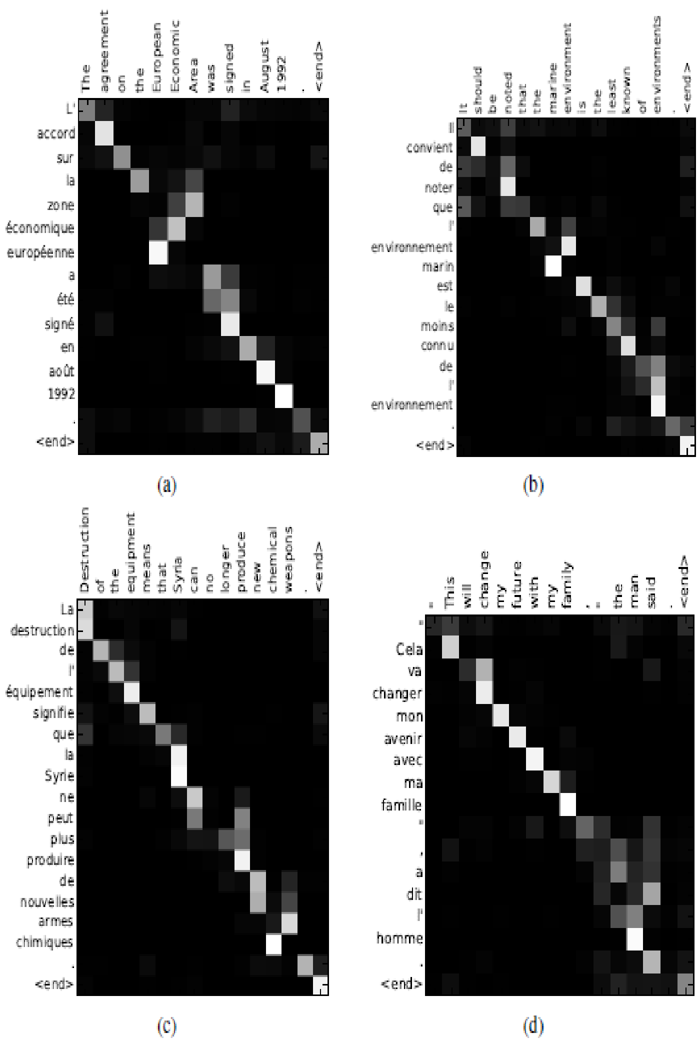
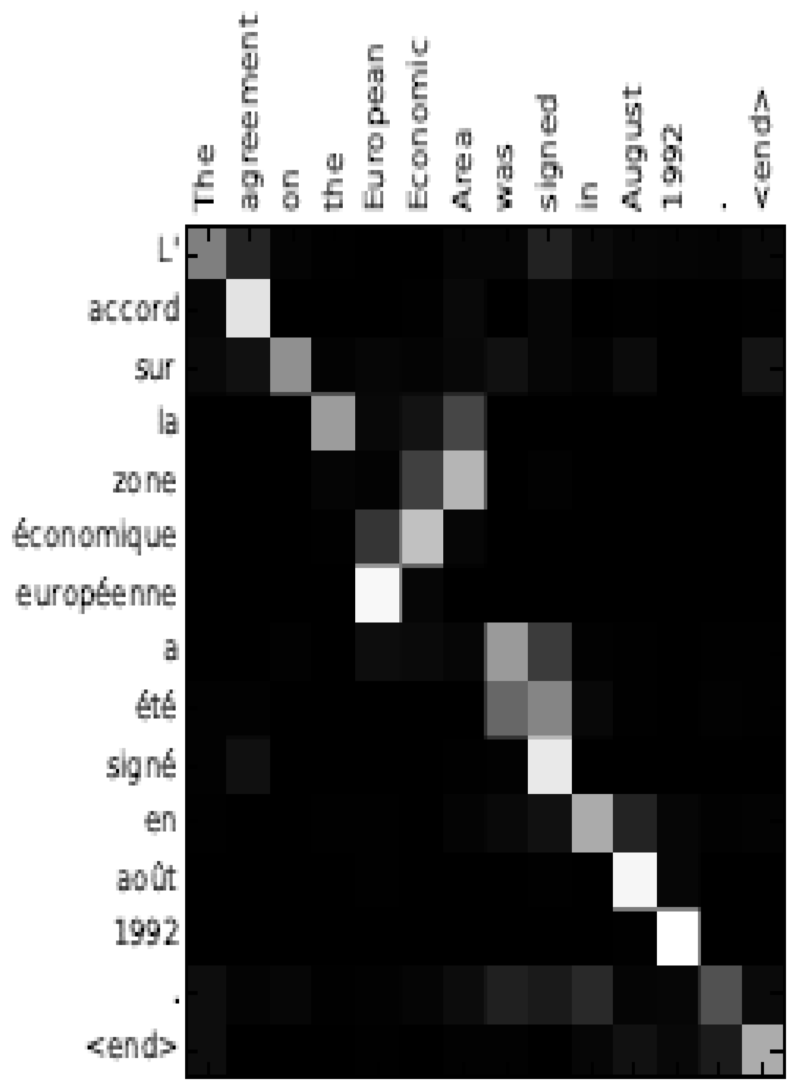
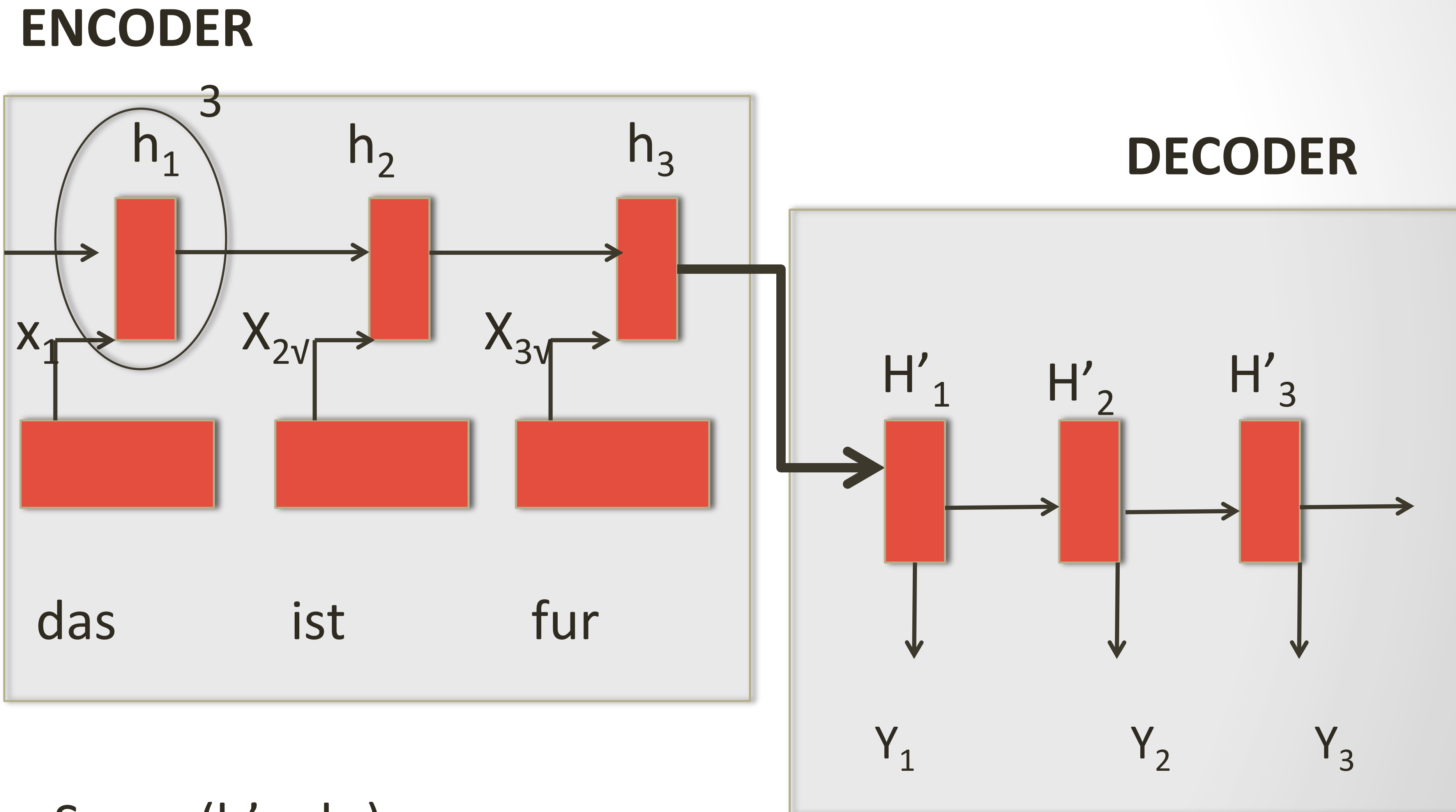


Figure 3: Four sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j-th source word for the i-th target word (see Eq. (6)), in grayscale (0: black, 1: white). (a) an arbitrary sentence. (b–d) three randomly selected samples among the sentences without any unknown words and of length between 10 and 20 words from the test set.



(a)

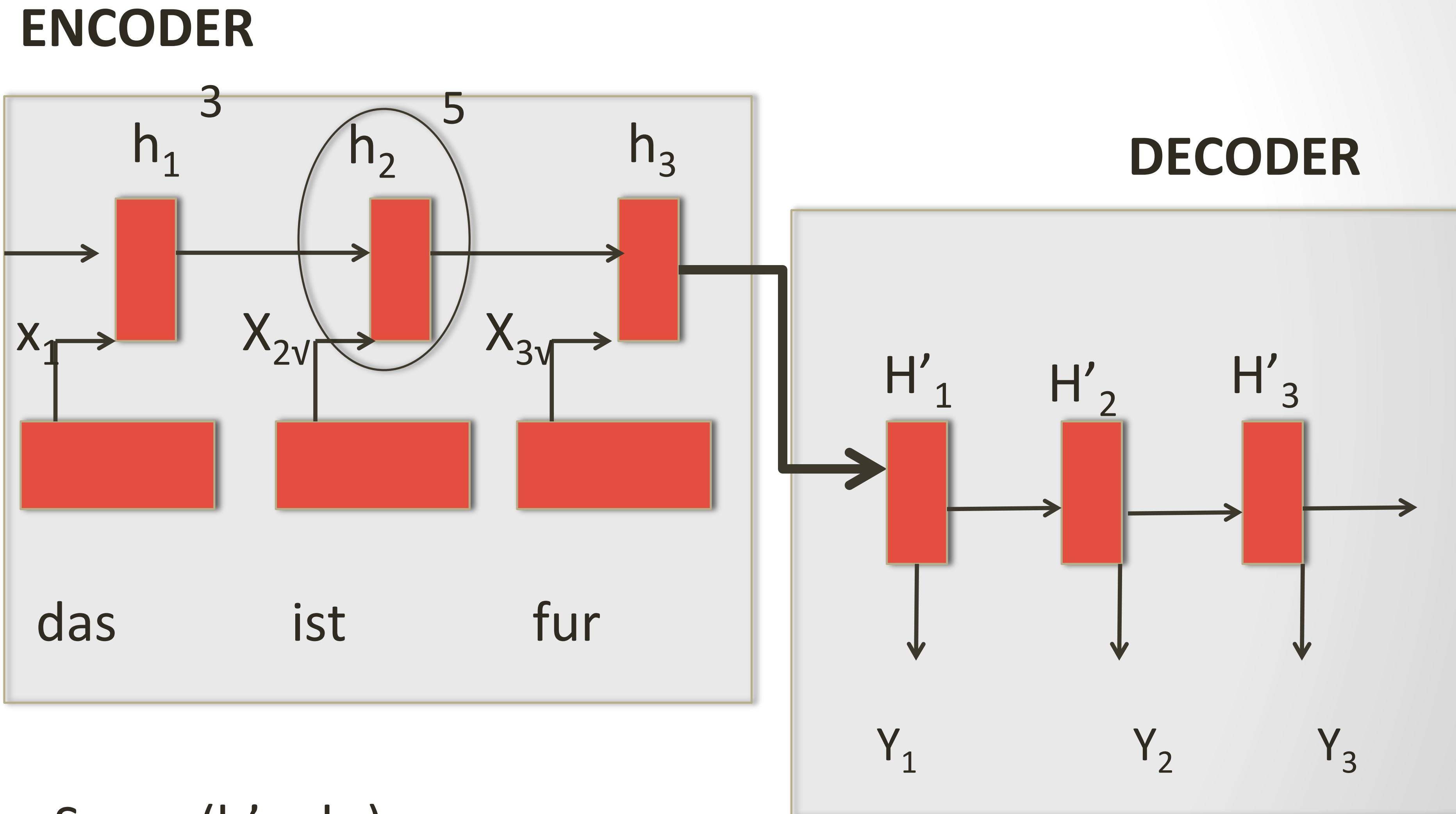
Attention Mechanism - Scoring



Score (h'_{t-1}, h_s)

That ?

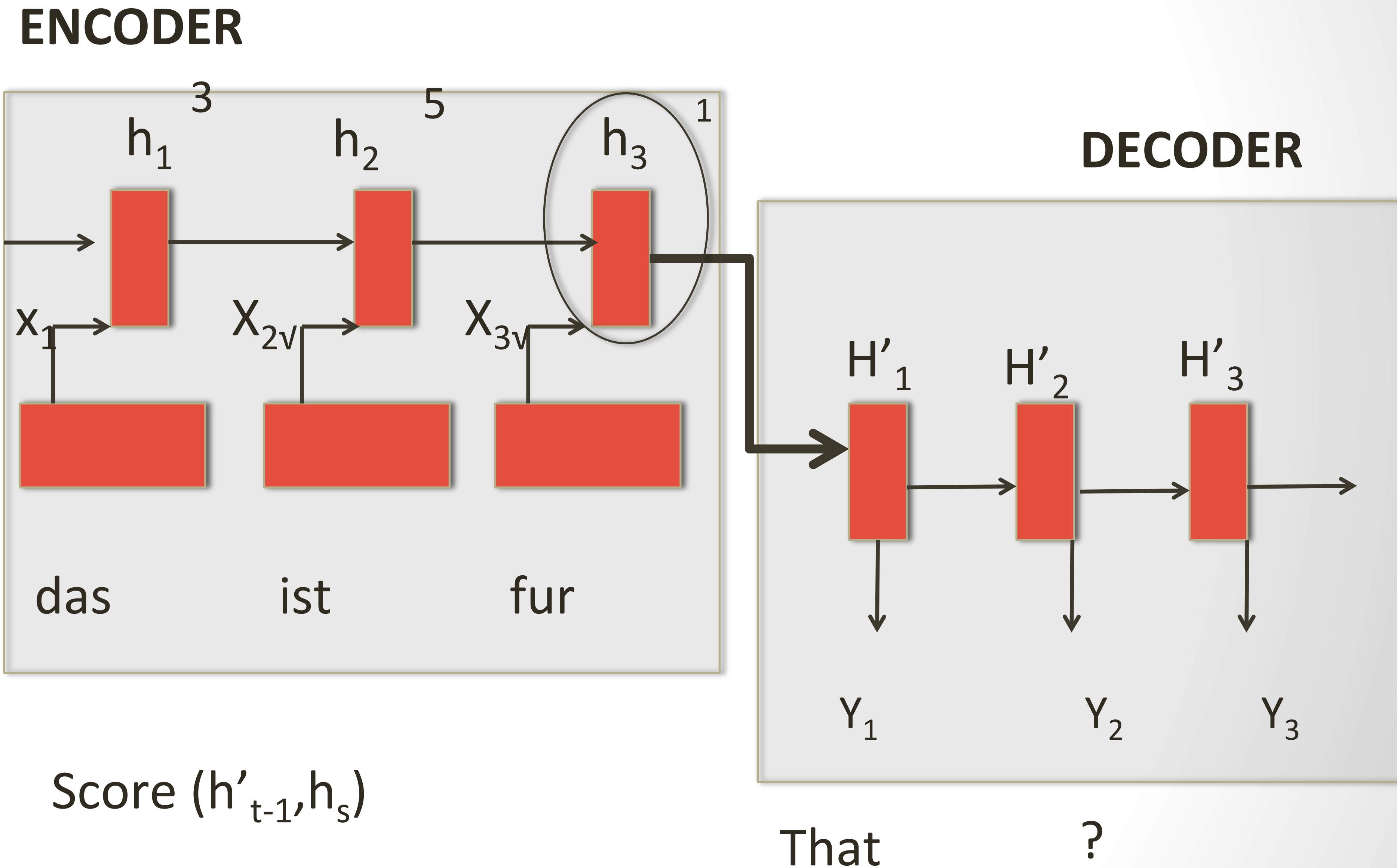
Attention Mechanism - Scoring



Score (h'_{t-1}, h_s)

That ?

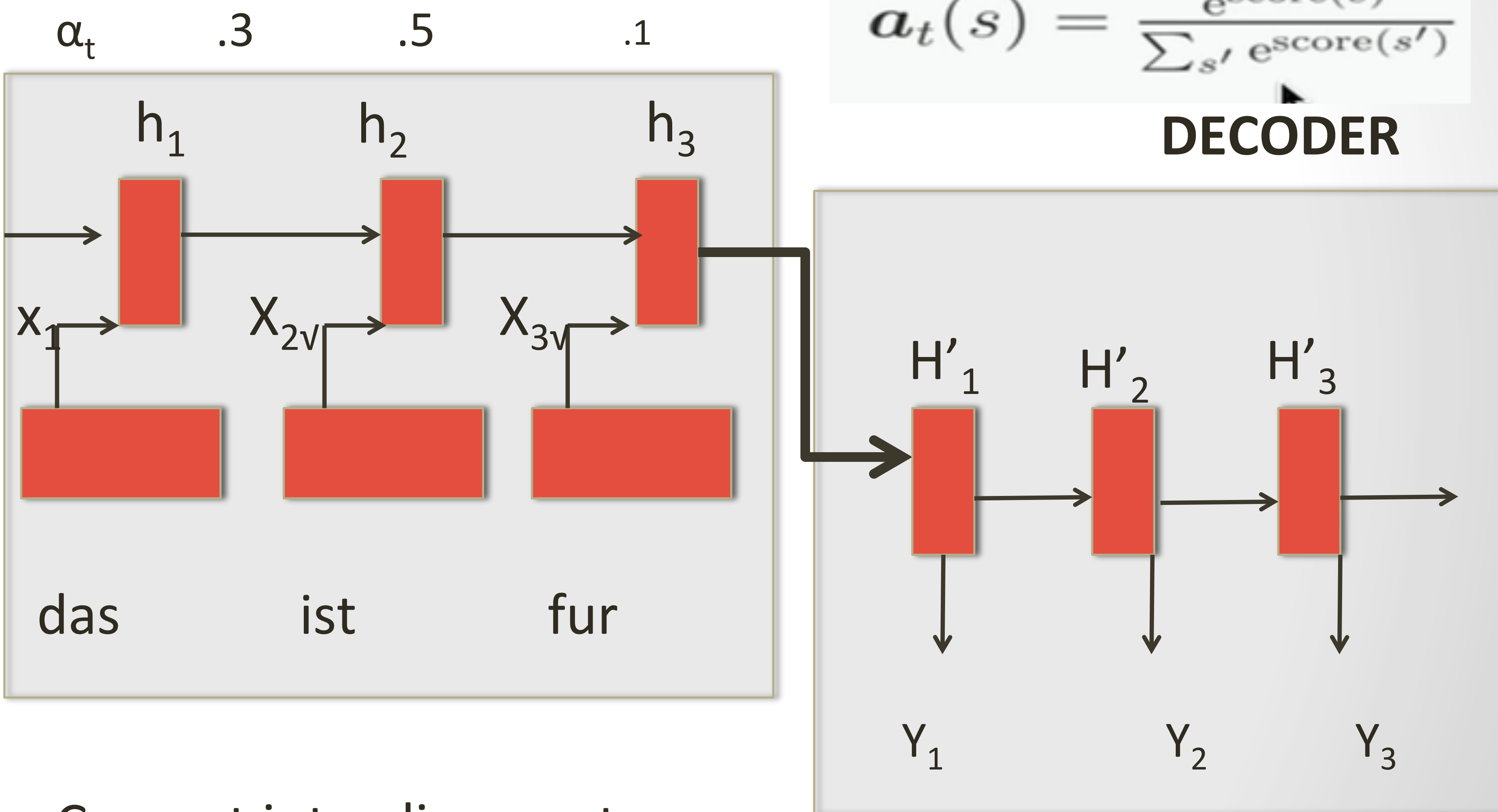
Attention Mechanism - Scoring



Score (h'_{t-1}, h_s)

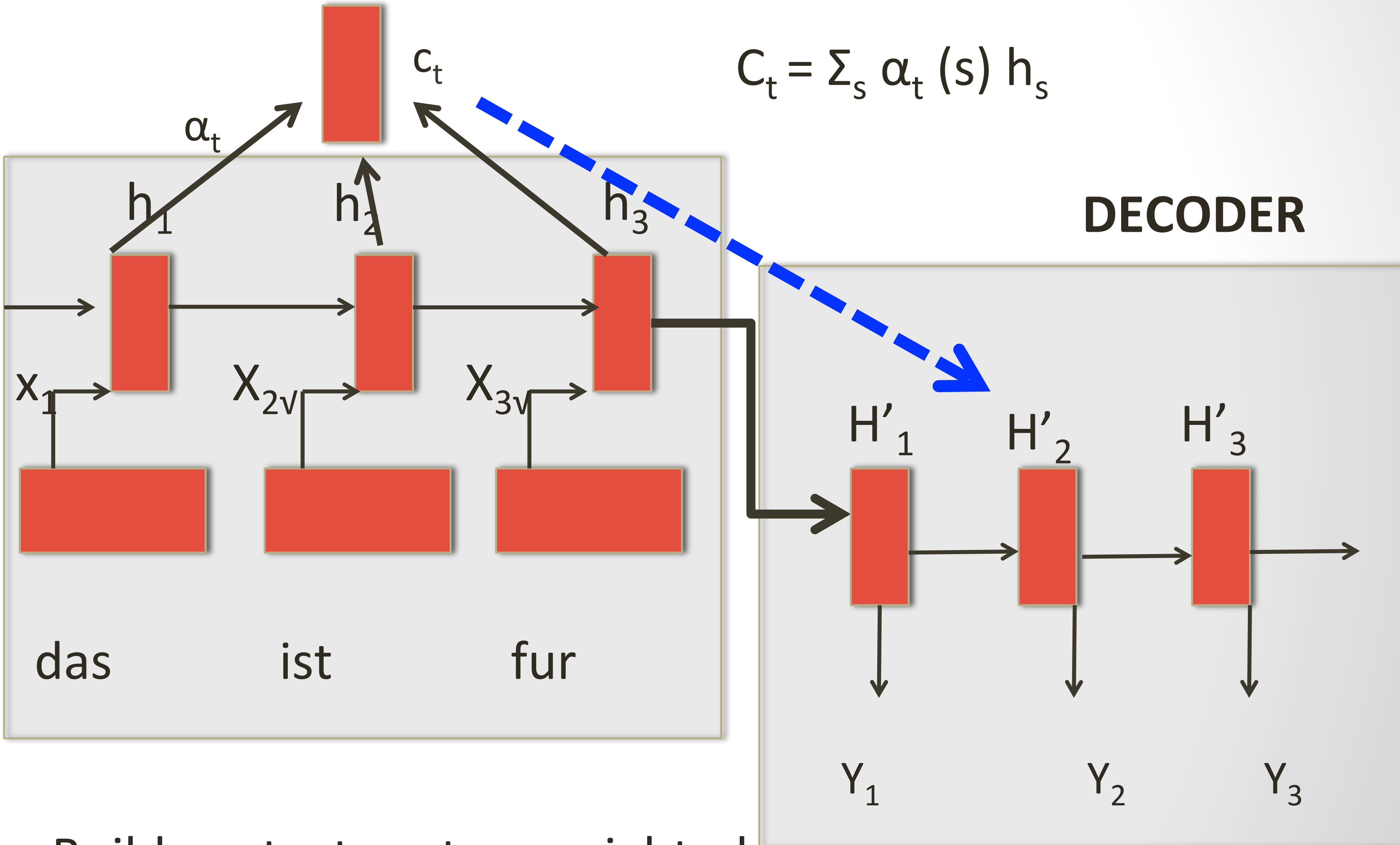
Attention Mechanism - Scoring

$$a_t(s) = \frac{e^{\mathrm{score}(s)}}{\sum_{s'} e^{\mathrm{score}(s')}}$$



Convert into alignment weights

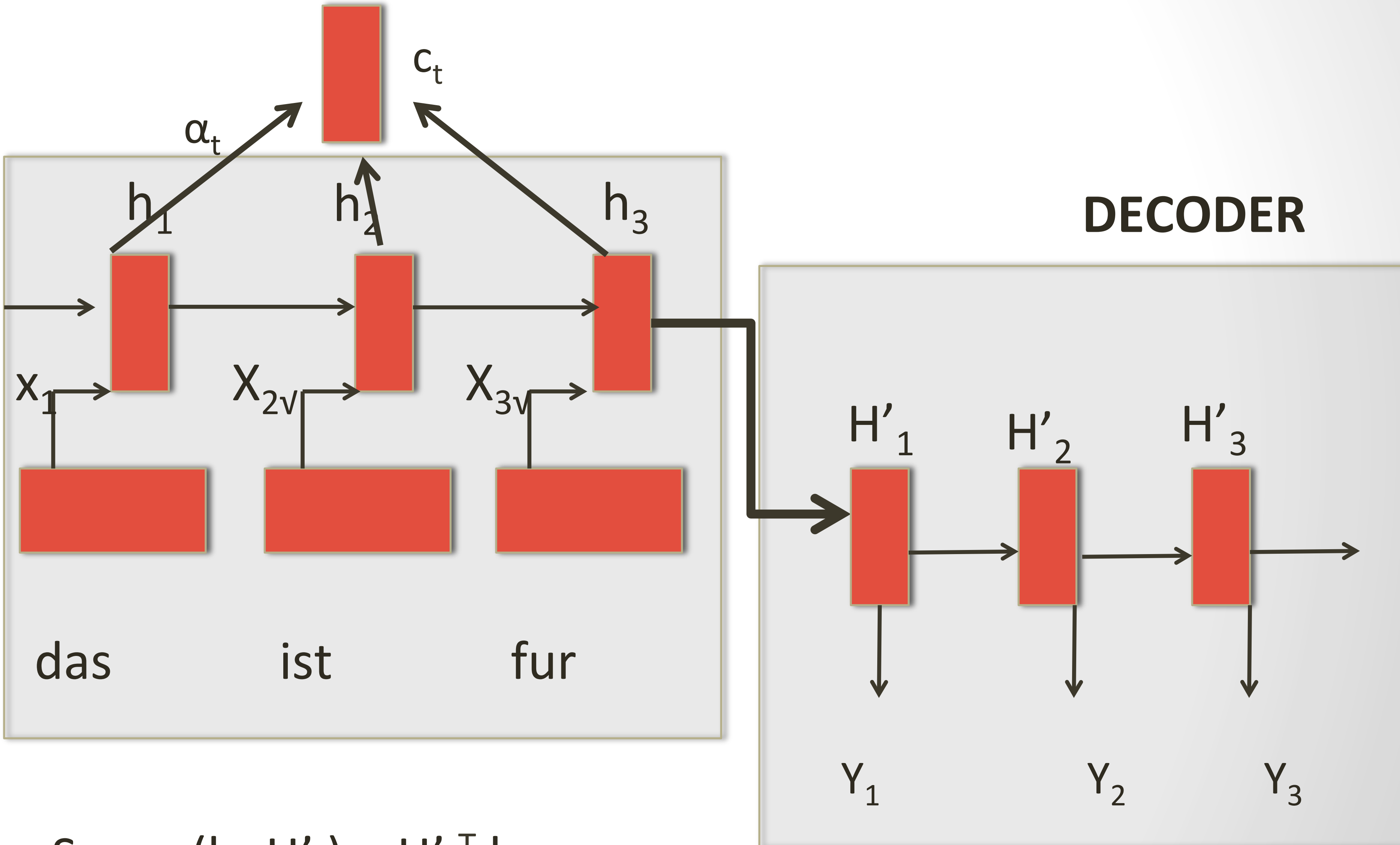
Attention Mechanism - Scoring



$C_t = \Sigma_s \, \alpha_t(s) \, h_s$

Build context vector: weighted average

How do you score it?



Score $(h_s, H'_t) = H'^{T}_{t} h_s$

or $= H'^{T}_{t} W_\alpha h_s$ (Luong et al 2015) ?

Performance

- Without attention, LSTM works quite well until a sentence gets longer than 30 words
- Attention does better, however, even with shorter sentences
- Other tricks in WMT 2017:
  - Improvements of 1.5 – 3 blue points (Edin)
  - Layer normalization, deeper networks (encoder depth of 5, decoder depth of 8)
  - Base Phrase Encodings (BPE)
    - Reduced vocabulary improves memory efficiency
  - Data: parallel, back-translated, duplicated monolingual

Questions?

Information Extraction

- Extraction of concrete facts from text
- Named entities, relations, events
- Often used to create a structured knowledge base of facts

- Kathy McKeown, a professor from Columbia University in New York City, took a train yesterday to Washington DC.

Named Entities

- Kathy McKeown$_{\text{per}}$, a professor from Columbia University$_{\text{org}}$ in New York City$_{\text{loc}}$, took a train yesterday to Washington DC$_{\text{loc}}$.

Named Entities, Relations

- Kathy McKeown$_{per}$, a professor from Columbia University$_{org}$ in New York City$_{loc}$, took a train yesterday to Washington DC$_{loc}$.

- Kathy McKeown from Columbia
- Columbia in New York City

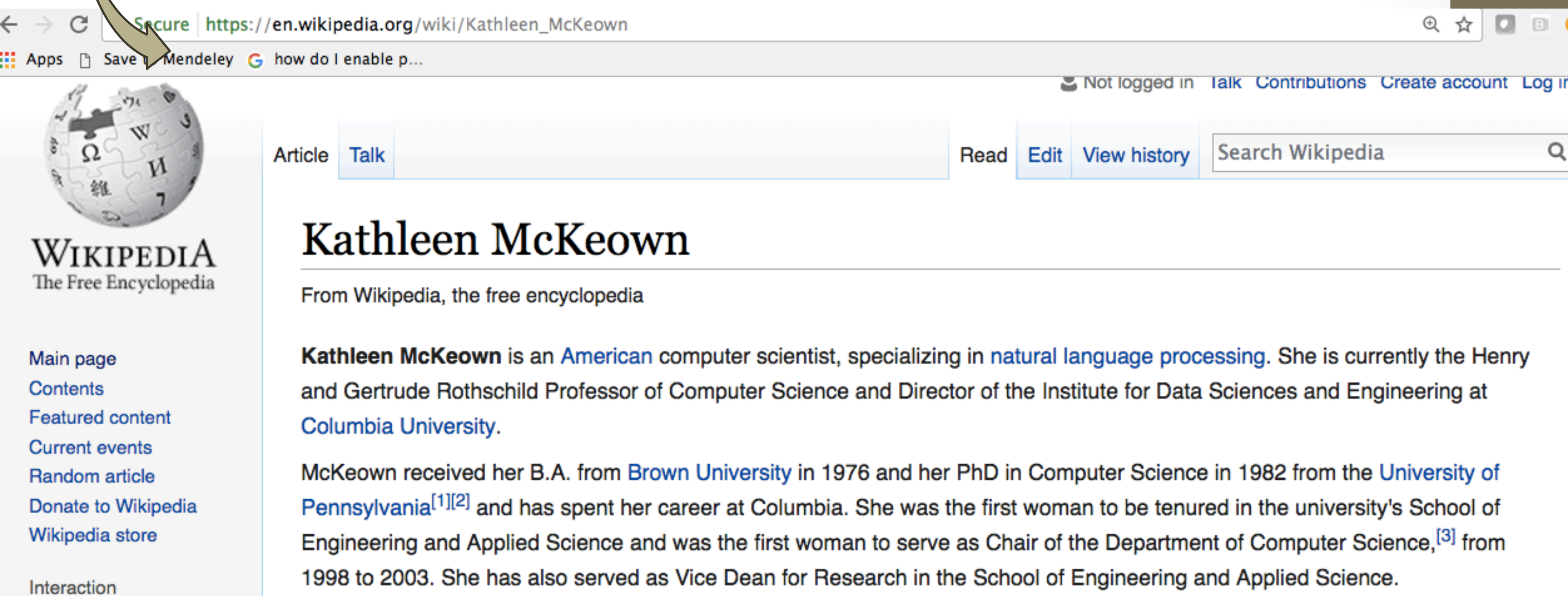
Named Entities, Relations, Events

- Kathy McKeown$_{per}$, a professor from Columbia University$_{org}$ in New York City$_{loc}$, took a train yesterday to Washington DC$_{loc}$.

- Kathy McKeown took a train (yesterday)

Entity Discovery and Linking

Kathy McKeown, a professor from Columbia University in New York City, took a train yesterday to Washington DC.


Secure https://en.wikipedia.org/wiki/Kathleen_McKeown

Apps Save to Mendeley how do I enable p...

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Kathleen McKeown

From Wikipedia, the free encyclopedia

Kathleen McKeown is an American computer scientist, specializing in natural language processing. She is currently the Henry and Gertrude Rothschild Professor of Computer Science and Director of the Institute for Data Sciences and Engineering at Columbia University.

McKeown received her B.A. from Brown University in 1976 and her PhD in Computer Science in 1982 from the University of Pennsylvania[1][2] and has spent her career at Columbia. She was the first woman to be tenured in the university's School of Engineering and Applied Science and was the first woman to serve as Chair of the Department of Computer Science,[3] from 1998 to 2003. She has also served as Vice Dean for Research in the School of Engineering and Applied Science.

State of the Art (English)

	F-measure
Named Entities (news)	89%
Relations (slot filling)	59%
Events (nuggets)	63%

Methods: Sequence labeling (MEMM, CRF), neural nets, distant learning

Features: linguistic features, similarity, popularity, gazeteers, ontologies, verb triggers

Where Have You Been Entity Discovery and Linking?

Grow with DEFT	2006-2011	2012-2017 *HENG JI, RPI*
Mention Extraction	Human (most)	Automatic
NIL Clustering	None	64 methods
Foreign Languages	Chinese (5%-10% lower than English)	**System for 282 languages (Chinese/Spanish comparable to/Outperform English); research toward 3,000 languages**
Document Size	-	500 →90,000 documents
Genre	News, web blog	**News, Discussion Forum, Web blog, Tweets**
Entity Types	PER, GPE, ORG	**PER, GPE, ORG, LOC, FAC, hundreds of fine-grained types for typing**
Mention Types	Name or all concepts (most)	Name, Nominal, Pronoun (for BeST)
KB	Wikipedia	Freebase → List only
Training Data	20,000 queries (entity mentions)	**500 → 0 documents; unsupervised linking comparable to supervised linking**
#(Good) Papers	62	110 (new KBP track at ACL); 6 tutorials at top conferences

Slide from Heng Ji

On the Horizon: Entity Discovery and Linking

Panel: Hoa Trang Dang, Jason Duncan, Heng Ji, Kevin Knight, Christopher Manning, Dan Roth

- Am going crazy
  - 3,000 languages
  - 10,000 entity types
  - All mention types
  - Multi-media
  - Streaming mode
  - List-only KB
  - Context-aware, living
  - No more training data
  - On-call evaluation
  - More non-traditional knowledge resources
  - Lots of dev and test sets in lots of languages

- Am staying cool
  - Success in end-to-end cold-start KBP
  - What's still wrong with name tagging
  - Smarter collective inference
  - Resolution of true aliases
  - Resolution of handles used as entity mentions

Slide from Heng Ji