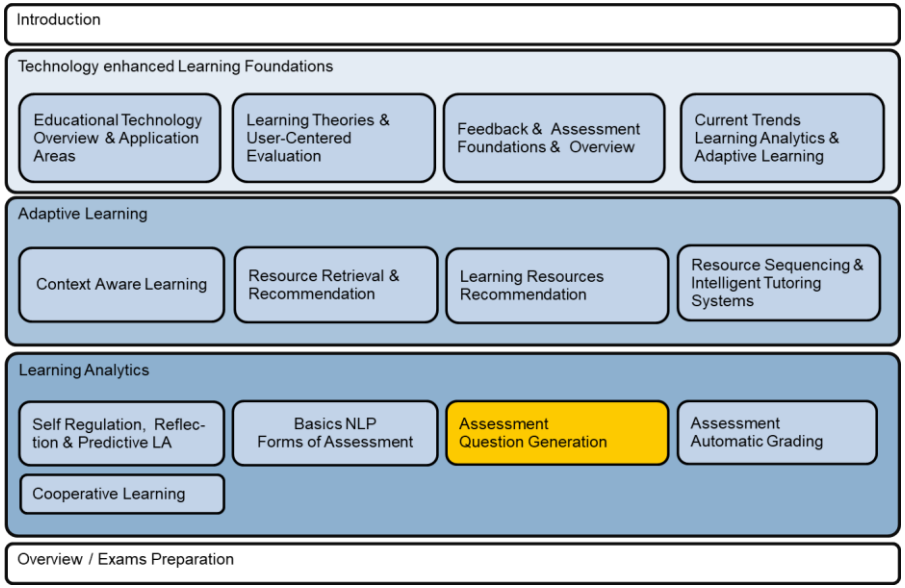
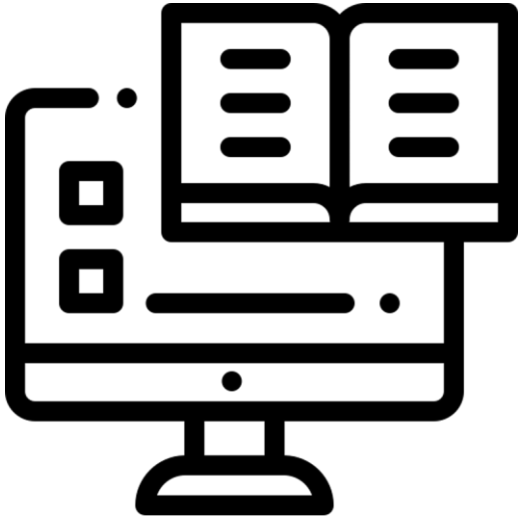
Assessment – Question Generation

Learning and Educational Technologies (SLKST)
22.01.2020

TECHNISCHE UNIVERSITÄT DARMSTADT





Icon made by https://www.flaticon.com/authors/freepik from www.flaticon.com

Tim Steuer
PD Dr.-Ing. **Christoph Rensing**
KOM - Multimedia Communications Lab

Lecture 11

TECHNISCHE UNIVERSITÄT DARMSTADT

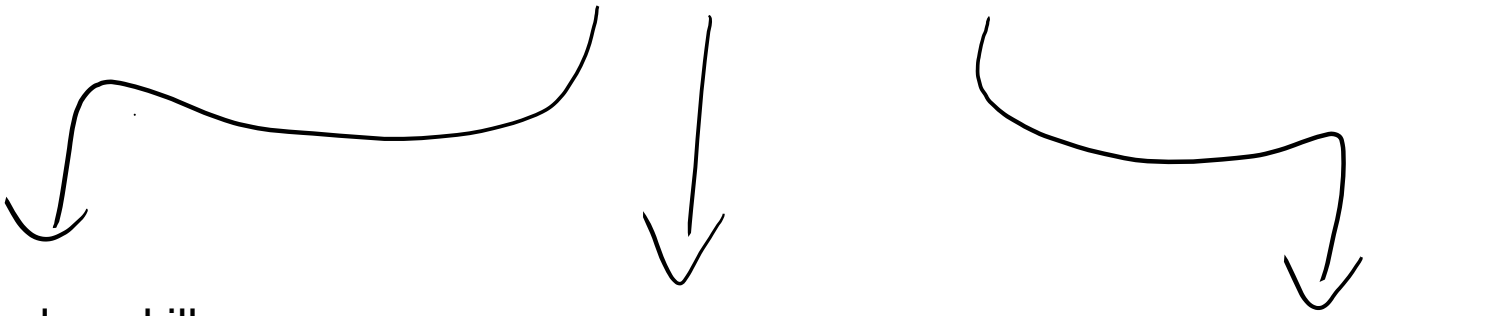
- Why Asking Questions is Important
- Automatic Question Generation Challenges in Education
- Evaluation

WHY ASKING QUESTIONS IS IMPORTANT

Questions in Education

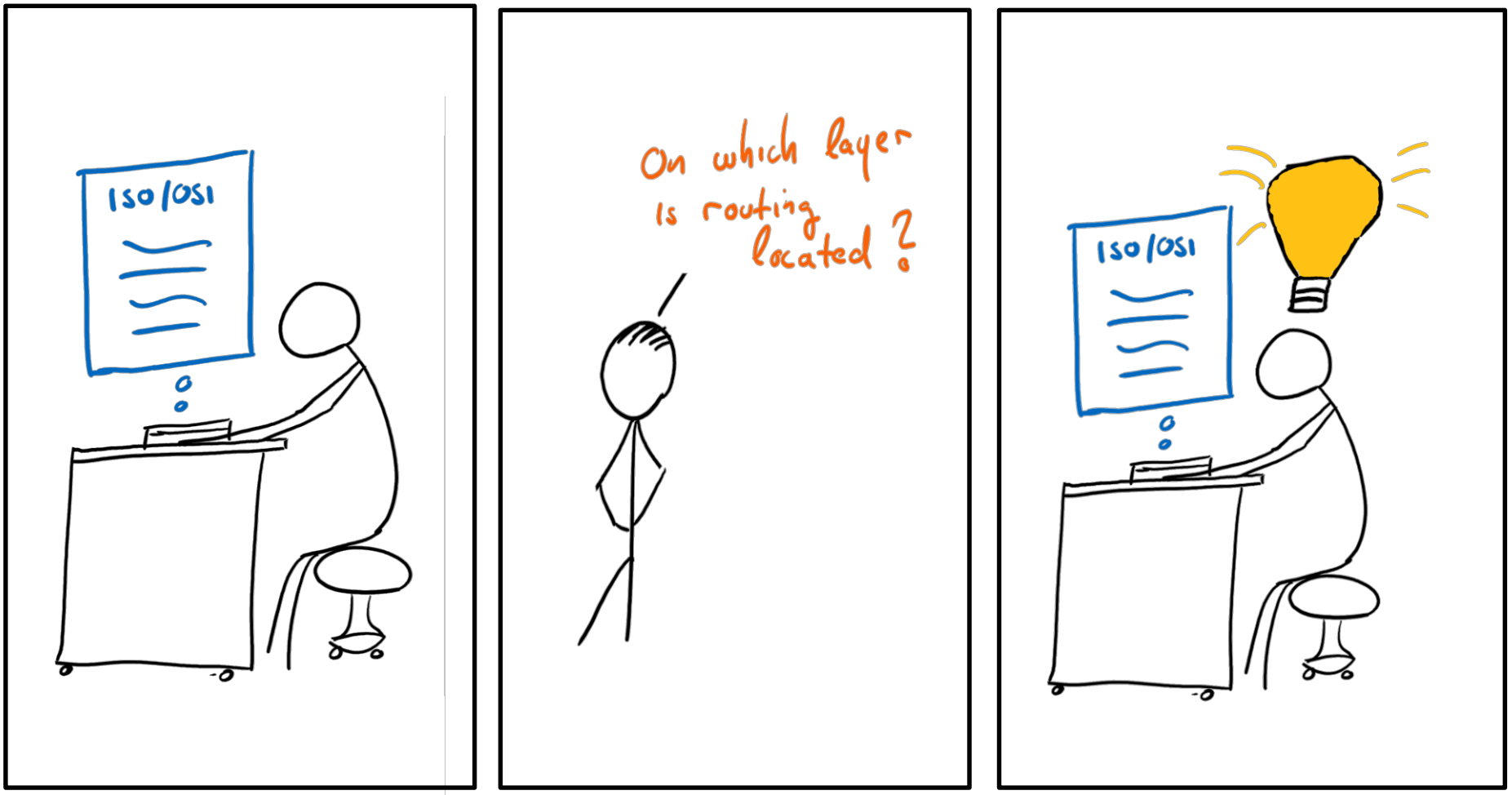
TECHNISCHE UNIVERSITÄT DARMSTADT

Why do we ask (automatic) questions when teaching? [1]



Attention: most systems mix categories

Why To Ask: Knowledge Acquisition



purpose: process information actively

benefits: Student

Example

Velociraptor is a genus of dromaeosaurid theropod dinosaur that lived approximately 75 to 71 million years ago during the latter part of the Cretaceous Period.[2] Two species are currently recognized, although others have been assigned in the past. The type species is V. mongoliensis; fossils of this species have been discovered in Mongolia. A second species, V. osmolskae, was named in 2008 for skull material from Inner Mongolia, China.

When did Velociraptors live?

They have been living in ____

...

Evidence: Adjunct Reading Questions

TECHNISCHE UNIVERSITÄT DARMSTADT

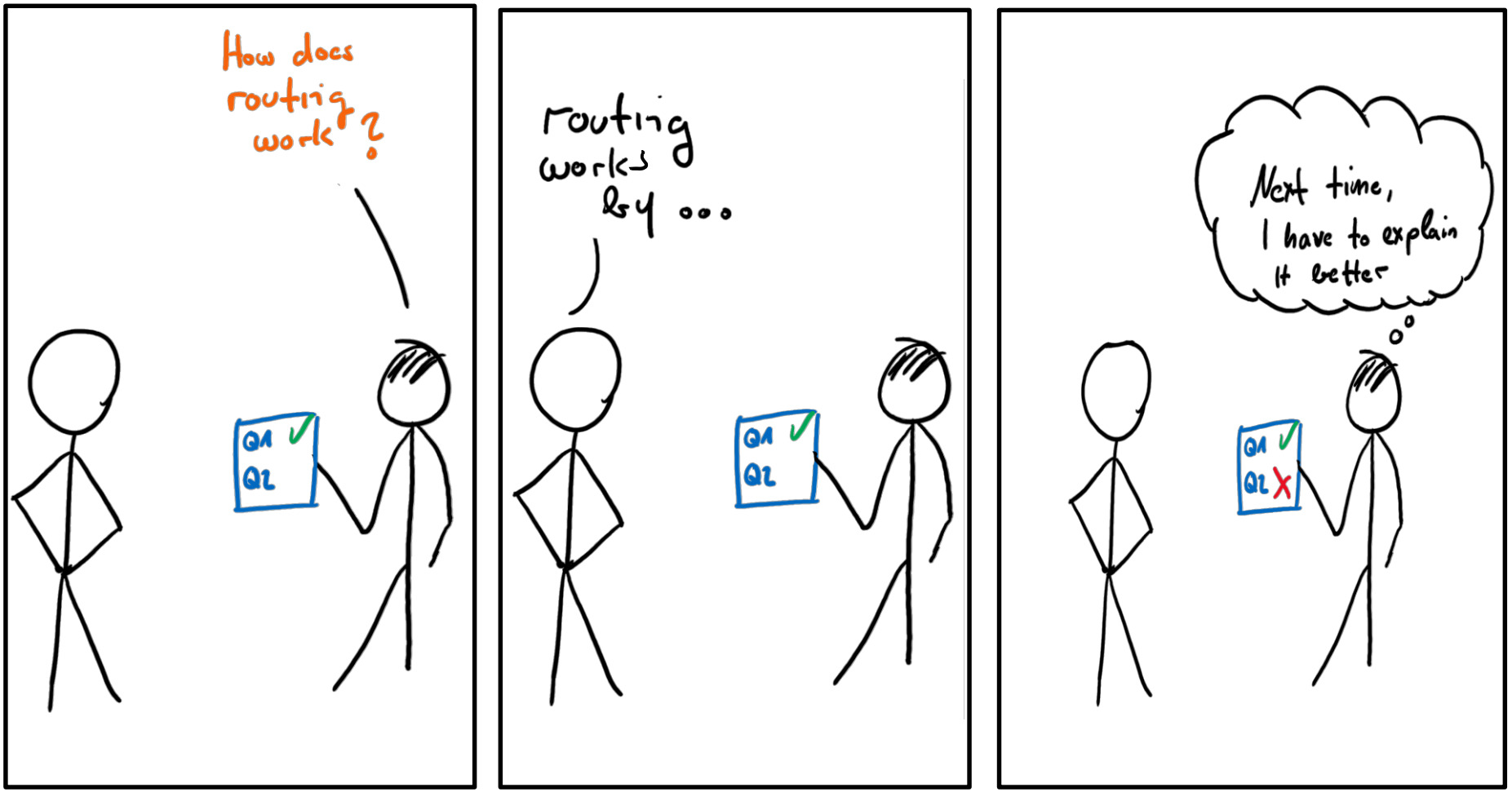
Anderson & Biddle 1975 [2] **(and many more):**

- Review of papers (N > 100)
- Evidence on various dimensions
  (when, how, what to ask…)

Main findings on effectiveness of questions

- verbatim questions → remember facts
- comprehension questions → higher level knowledge
- only verbatim → might hinders deeper comprehension

Why To Ask: Formative Knowledge Assessment



purpose: determining knowledge of learner

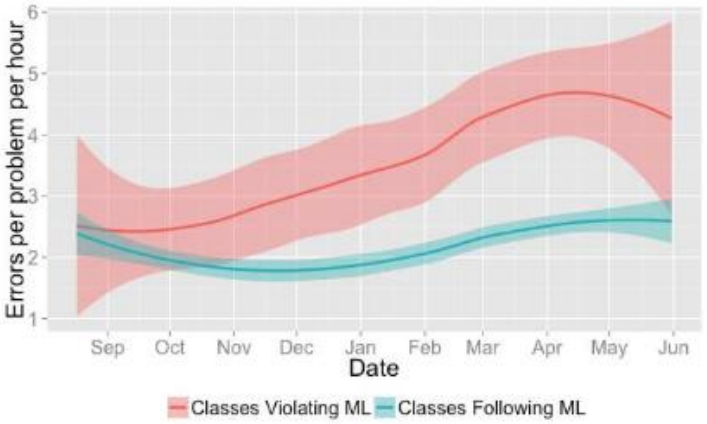
benefits: Teacher / Educational technology

Usefulness of (formative) Assessment

TECHNISCHE UNIVERSITÄT DARMSTADT

Formative assessment determines prior knowledge

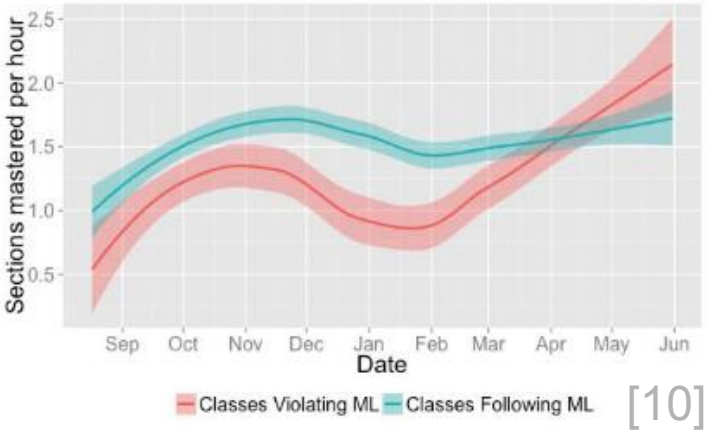
- good instruction → prior knowledge
  - Subsumption Theory (Ausubel; [3])
  - Mastery Learning (Bloom; [4])
  - …



(a)

Formative assessments is basis for continuous feedback

- tremendous effects on learning



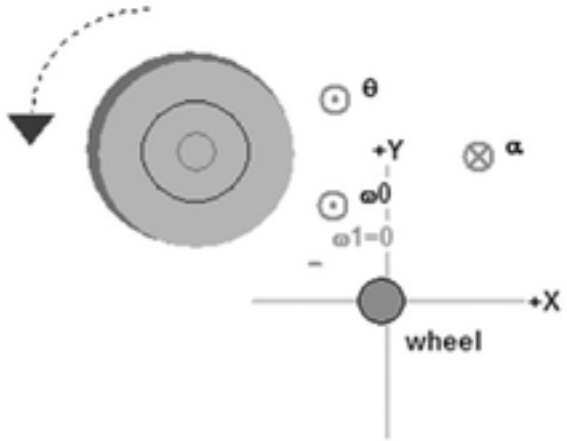
[10]

Example: Andes Tutoring System

Answer: 20 s

Through how many radians does the wheel turn before it comes to a complete stop?

Answer: 100 rad



√ x	axis	θx=0°			
√ α	magnitude of the average Angular...	ωα=180°	α_x	α_y	α_z
√ θ	magnitude o				
√ ω0	magnitude o				
√ ω1	magnitude of the instantaneous A...		ω1_x	ω1_y	ω1_
√ t	duration of time from T0 to T1				

1. ω0_z=10 rad/s
2. α_z = -0.5 rad/s^2
3. ω1_z= 0
4. ω1_z = ω0_z + α_z*t
5. t = 20 s
6. θ_z = ω0_z*t + 0.5*α_z*t^2
7. θ_z = 100 rad
8.
9.
10.
11.
12.
13.
14.
15.

T: Units are inconsistent.
OK

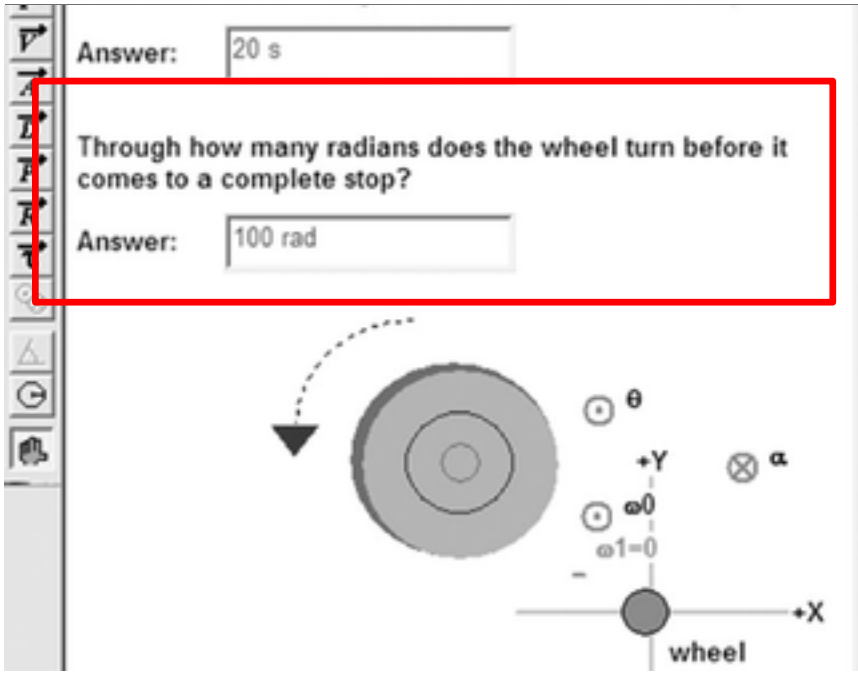
T: Undefined variable: ω01_z
Explain further OK

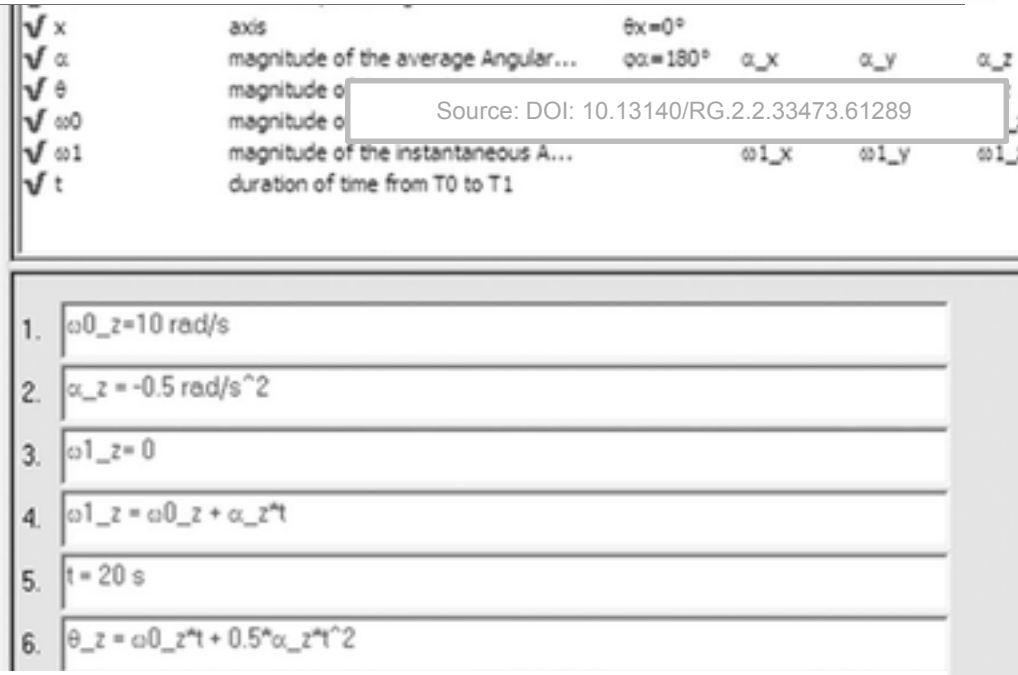
T: Units are inconsistent.
OK

T: Undefined variable: θ1_z
Explain further OK

T: Unable to solve for θ_z. Try the light bulb if you need a hint about a step that still needs to be done.

Example: Andes Tutoring System



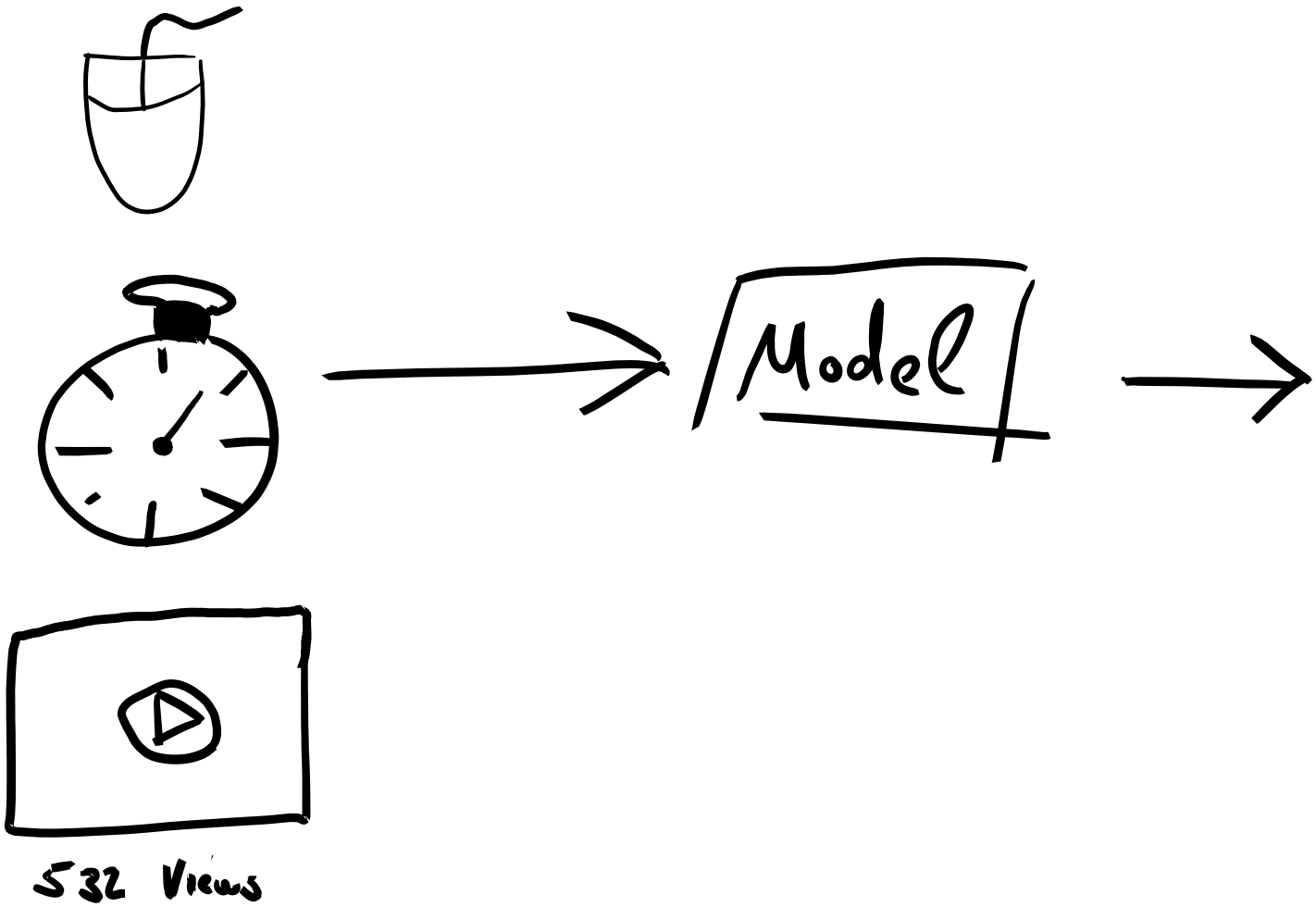


Andes Example Results [5]

Table 3: Results from hour exams evaluations

Year	1999	2000	2001	2002	2003
Number of Andes students	173	140	129	93	93
Number of control students	162	135	44	53	44
Andes mean (stan. dev)	73.7 (13.0)	70.0 (13.6)	71.8 (14.3)	68.2 (13.4)	71.5 (14.2)
Control mean (stan. dev)	70.4 (15.6)	57.1 (19.0)	64.4 (13.1)	62.1 (13.7)	61.7 (16.3)
P(Andes = Control)	0.036	< .0001	.003	0.005	0.0005
Effect size	0.21	0.92	0.52	0.44	0.60

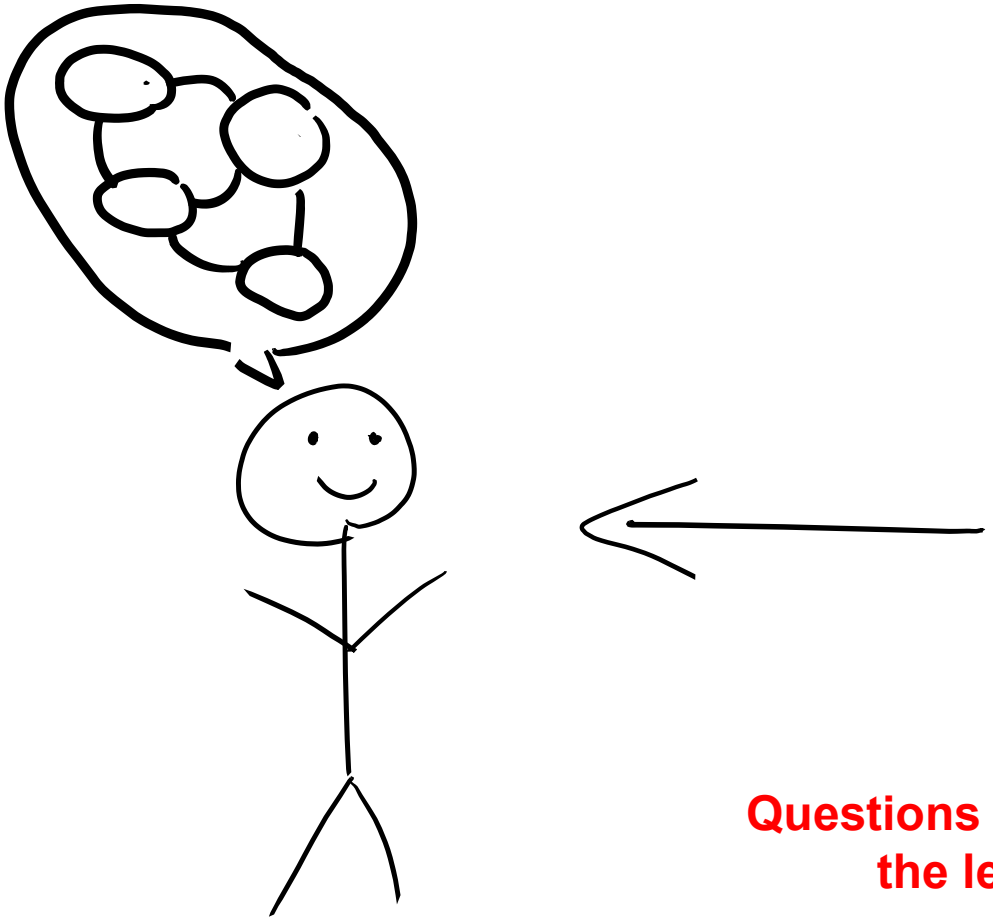
Besides Human Needs: Learning Analytics



Student knows "Networking"

Really?

Learning Analytics: How Teachers do it



"What do you know about network switches?"

Questions provide valuable insights in the learners understanding

Why To Ask: Socratic Dialog

TECHNISCHE UNIVERSITÄT DARMSTADT

Interaction Hypothesis:

People tend to learn better when all of the following apply

One-to-one tutoring situation

Interaction with the material (via natural language)

Natural Language dialogs are used

Example: Auto Tutor

Source: DOI: 10.1007/s40593-014-0029-5

Interaction Hypothesis: Evidence

TECHNISCHE UNIVERSITÄT DARMSTADT

VanLehn et. al. 2007: When are tutorial dialogs more effective than reading? [6]

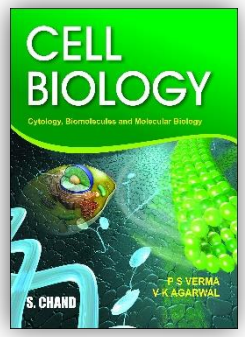
- Testing the Interaction Hypothesis
- Review of many papers + additional studies

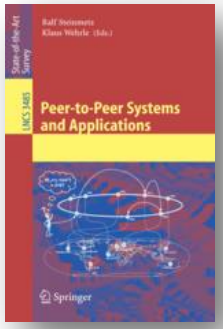
Main Findings

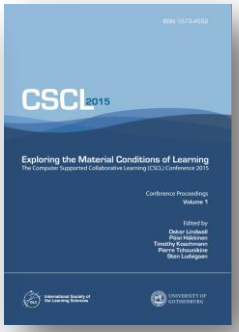
- (language) Interaction is more efficient than no interaction
- Novices need high-interaction (e.g. tutorial dialog)
- Intermediate students need no interaction or low-interaction (e.g. questioning)

AUTOMATIC QUESTION GENERATION METHODS FOR EDUCATION

A concrete Educational AQG Task Definition







Educational Books

Papers

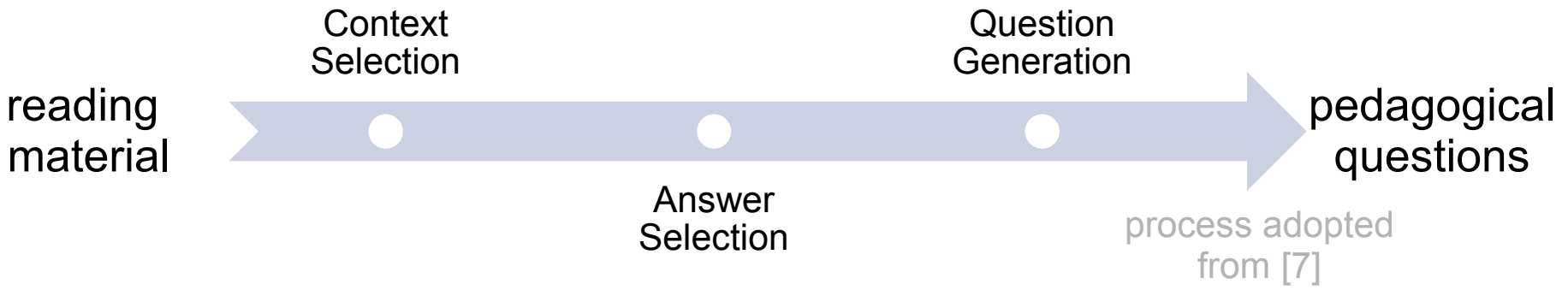
What is a web browser?

~~What is the authors name?~~

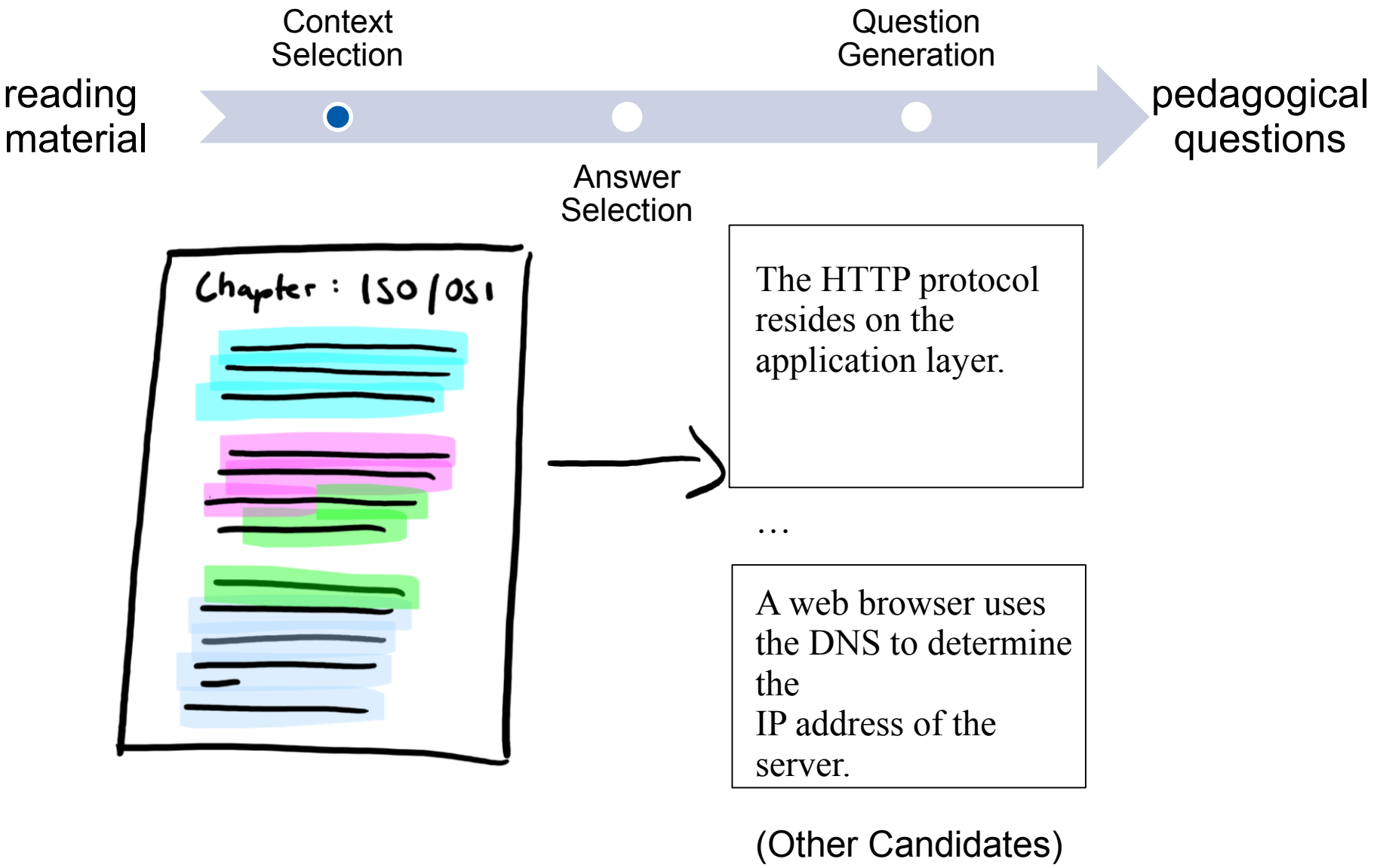
~~What are the advantages and disadvantages of peer-to-peer systems?~~

Factual texts → Educational AQG → pedagogical Questions (Bloom 1-2)

A possible AQG Process



Educational AQG: Context Selection



Context Selection Algorithmic Example

TECHNISCHE UNIVERSITÄT DARMSTADT

Input

- Long factual text

Output

- *N* Sentences describing the gist of the text

Example: LexRank Algorithm Family [8]

1. Represent text as vectors
2. Calculate similarity between every sentence pair
3. Construct a undirected graph based on similarities
4. Use graph properties to extract most important sentences

Context Selection Algorithmic Example

TECHNISCHE UNIVERSITÄT DARMSTADT

Input

- Long factual text

Output

- *N* Sentences describing the gist of the text

Example: LexRank Algorithm Family

1. Represent text as vectors ← What can you use there?
2. Calculate similarity between every sentence pair
3. Construct a undirected graph based on similarities
4. Use graph properties to extract most important sentences

Context Selection Algorithmic Example

TECHNISCHE UNIVERSITÄT DARMSTADT

Input

- Long factual text

Output

- *N* Sentences describing the gist of the text

Example: LexRank Algorithm Family

1. Represent text as vectors
2. Calculate similarity between every sentence pair ← What can you use there?
3. Construct a undirected graph based on similarities
4. Use graph properties to extract most important sentences

Context Selection Algorithmic Example

TECHNISCHE UNIVERSITÄT DARMSTADT

Input

- Long factual text

Output

- *N* Sentences describing the gist of the text

Example: LexRank Algorithm Family

1. Represent text as vectors
2. Calculate similarity between every sentence pair
3. Construct a undirected graph based on similarities

← How many nodes / edges has the graph?

4. Use graph properties to extract most important sentences

Context Selection Algorithmic Example

TECHNISCHE UNIVERSITÄT DARMSTADT

Input

- Long factual text

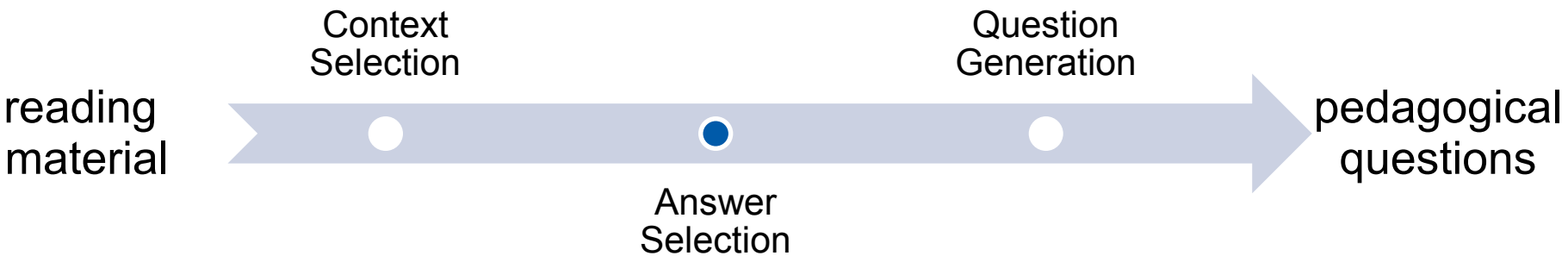
Output

- *N* Sentences describing the gist of the text

Example: LexRank Algorithm Family

1. Represent text as vectors
2. Calculate similarity between every sentence pair
3. Construct a undirected graph based on similarities
4. Use graph properties to extract most important sentences ← What can you use there?

Educational AQG: Answer Selection



The HTTP protocol resides on the application layer.

The HTTP protocol resides on the application layer.

Answer Selection Algorithmic Approaches

TECHNISCHE UNIVERSITÄT DARMSTADT

Input

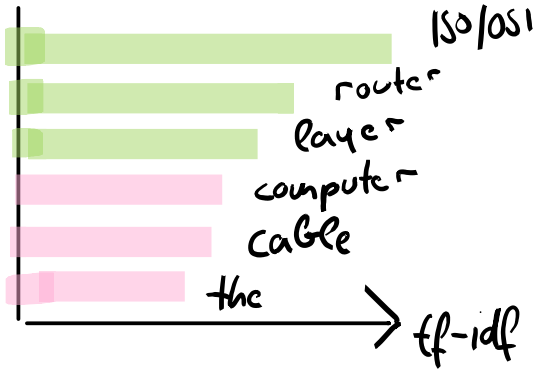
- Key Sentences
  (+ source corpus of the sentence)

Output

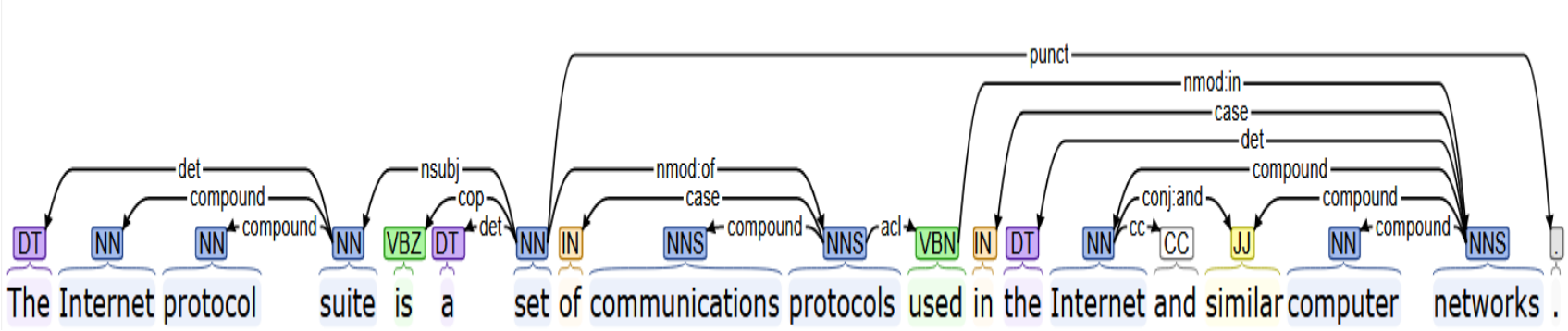
- Words in the sentences being a good answer for a question

Example Solutions

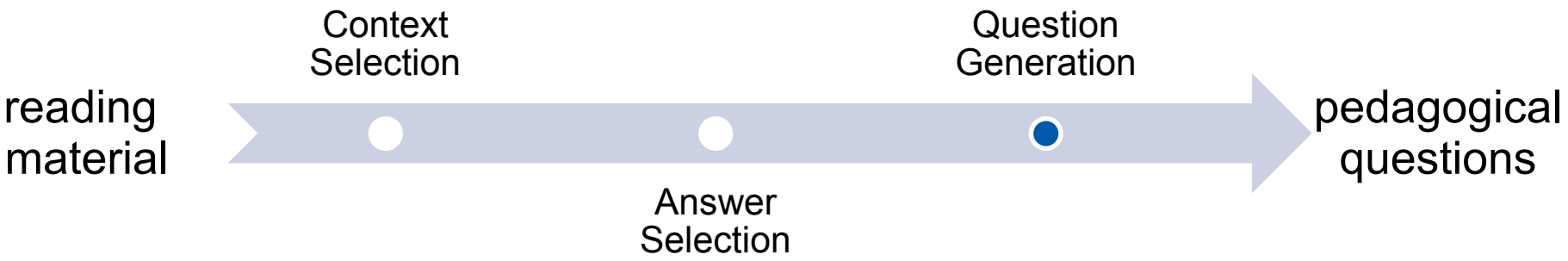
- TF-IDF based keywords filters



- Extraction via textual analysis:
  parse-trees or Part of Speech (POS)
  and Named entity recognition (NER)



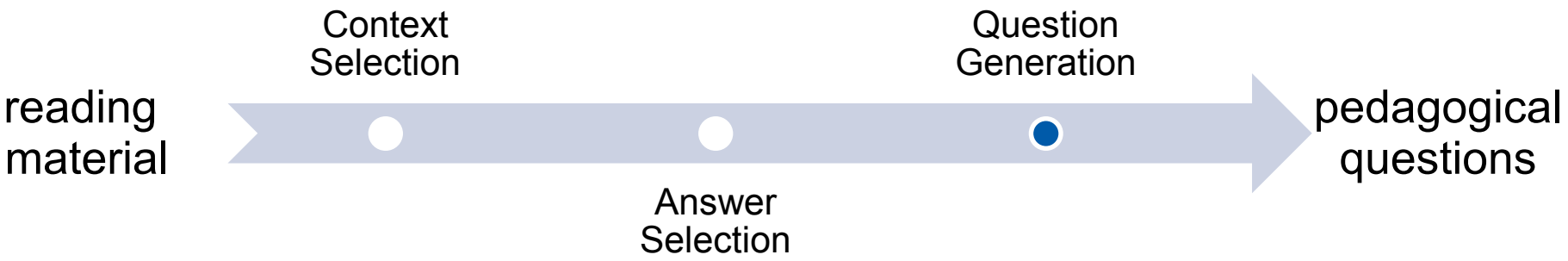
Educational AQG: Question Generation



The HTTP protocol resides on the application layer.

→ What protocol is on the application layer?

Educational AQG: Question Generation



The HTTP protocol resides on the application layer.

→ What protocol is on the application layer?

Methods

- Rule-based
- Templates
- Statistical

Statistical Question Generation (1)

TECHNISCHE UNIVERSITÄT DARMSTADT

Input

- Key Sentences + Answer
  (+ source corpus of the sentence)

Output

- Question asking for the marked answer

Basic Idea

To be or not to ⎵ ?

It is raining. Tim forgot his ⎵

Tim's drawings are simple.
⎵ are Tim's drawings?

Observation 1: The probability of the next word in a sentence is **dependent on the previous words**

Observation 2: The probability of the next word in a sentence is **dependent on the sentence context.**

Observation 3: The probability of words in a question **depends on the answer.**

Statistical Question Generation: Training

TECHNISCHE UNIVERSITÄT DARMSTADT

During training

learn $P(w_i = w | w_{i-1}, \dots, w_0, context, answer)$

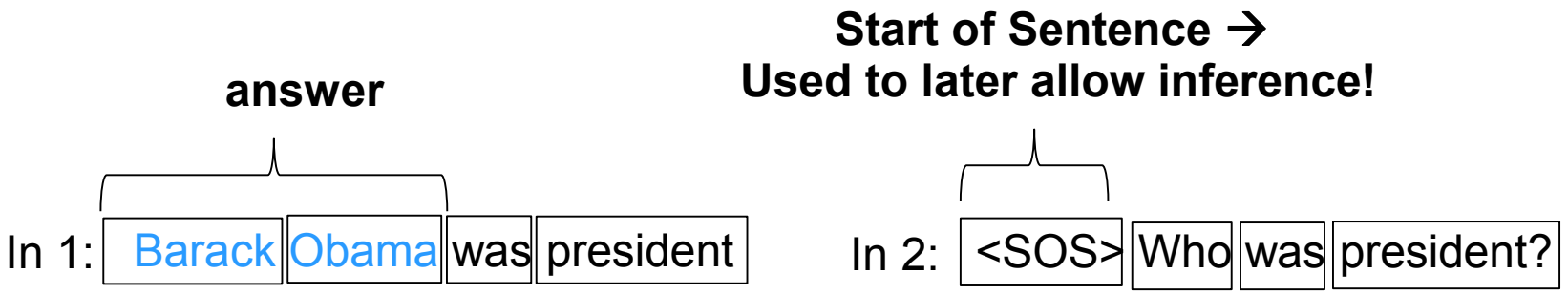
data

gold questions | answers | context sentences

Statistical Question Generation Training Example (Seq2Seq Model)

learn $P(w_i = w | w_{i-1}, \dots, w_0, context, answer)$

Gold: Who was president?



Statistical Question Generation(2)
Training Example

learn $P(w_i = w | w_{i-1}, \dots, w_0, context, answer)$

Gold: Who was president?

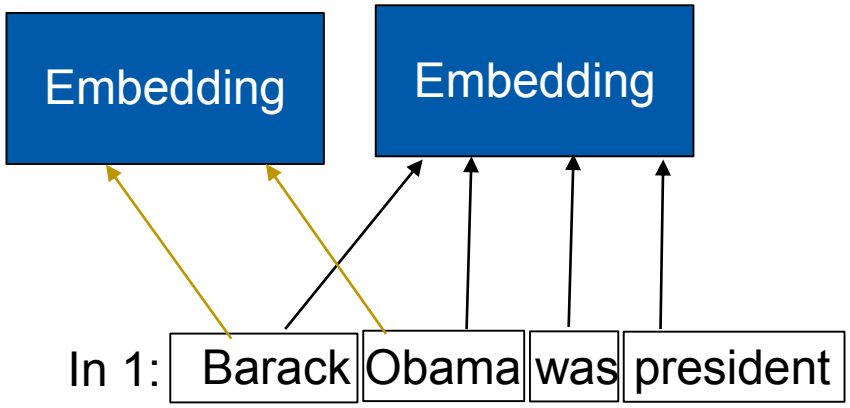
In 1: Barack | Obama | was | president

In 2: <SOS> | Who | was | president?

Statistical Question Generation(2)
Training Example

learn $P(w_i = w|w_{i-1}, \dots, w_0, context, answer)$

Gold: Who was president?

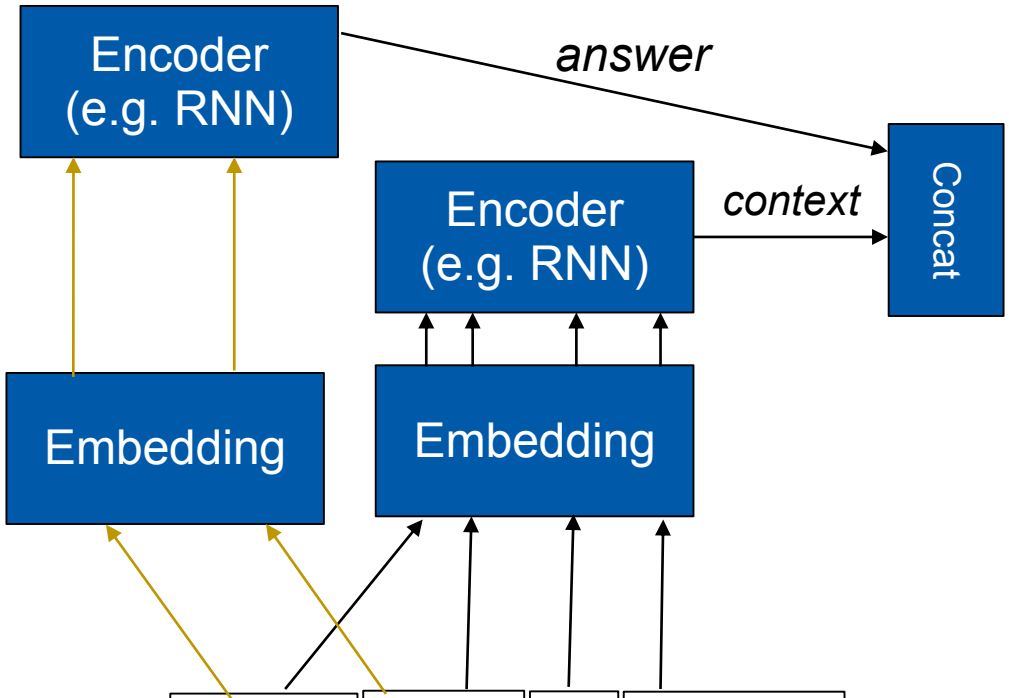


In 2: <SOS> Who was president?

Statistical Question Generation(2)
Training Example

learn $P(w_i = w | w_{i-1}, \dots, w_0, context, answer)$

Gold: Who was president?



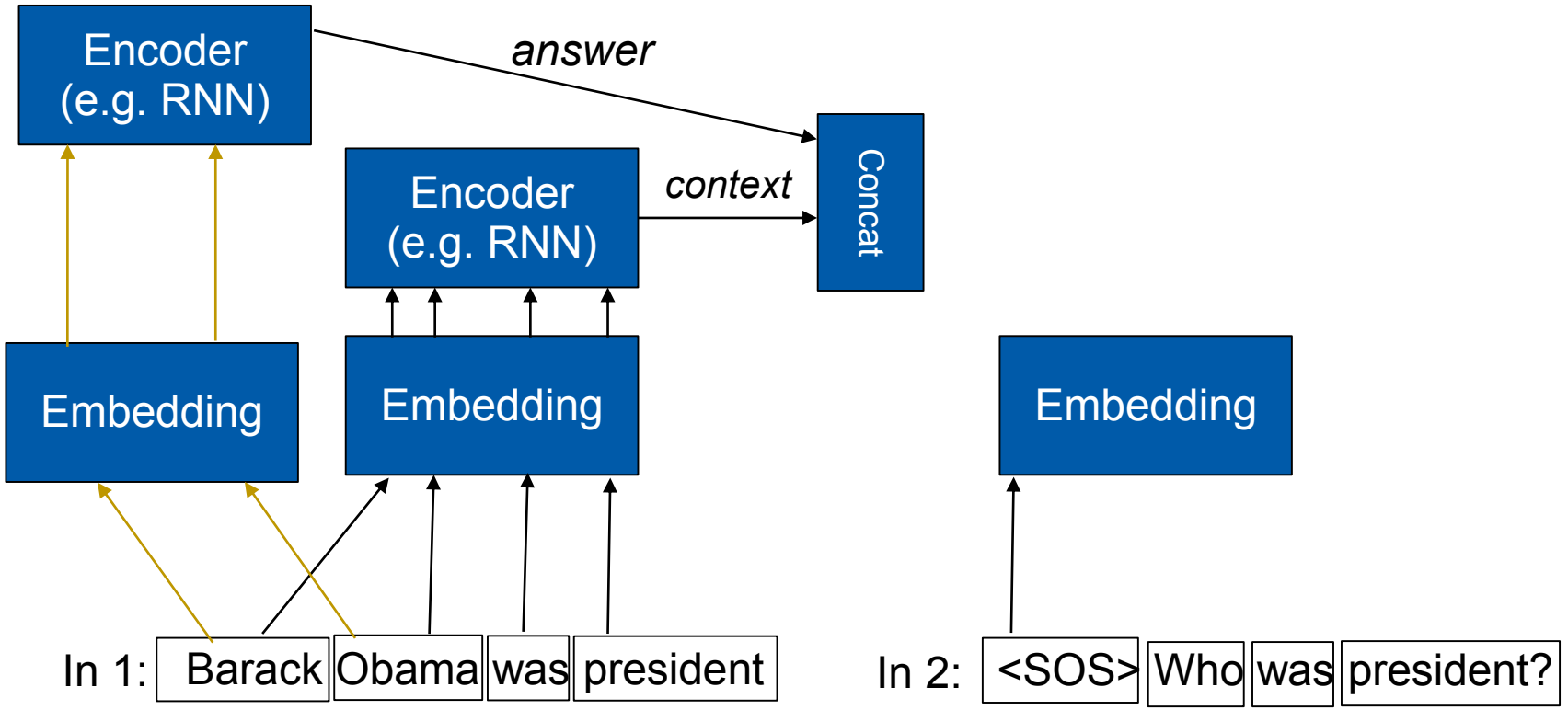
In 1: Barack | Obama | was | president

In 2: <SOS> | Who | was | president?

Statistical Question Generation(2)
Training Example

learn $P(w_i = w | w_{i-1}, \dots, w_0, context, answer)$

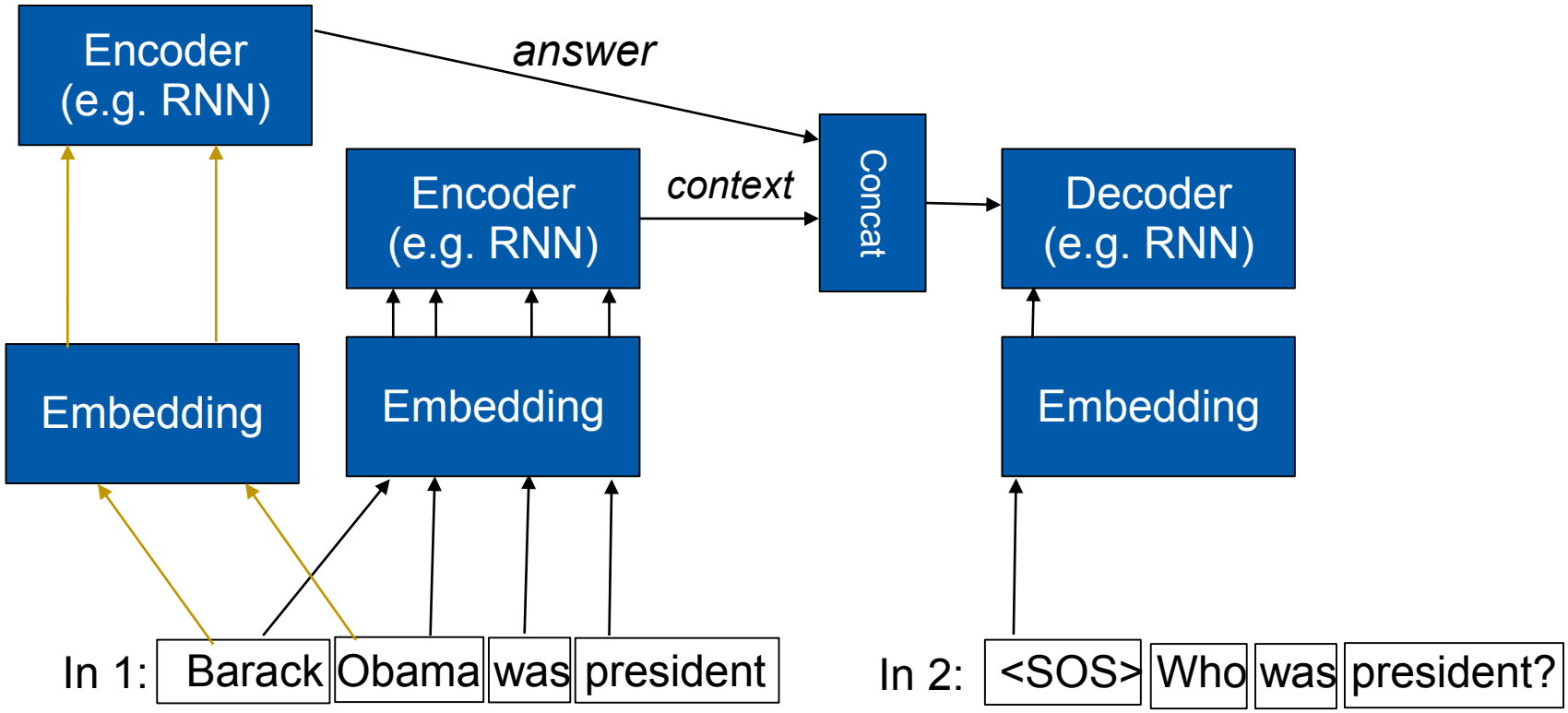
Gold: Who was president?



Statistical Question Generation(2)
Training Example

learn $P(w_i = w|w_{i-1}, \dots, w_0, context, answer)$

Gold: Who was president?

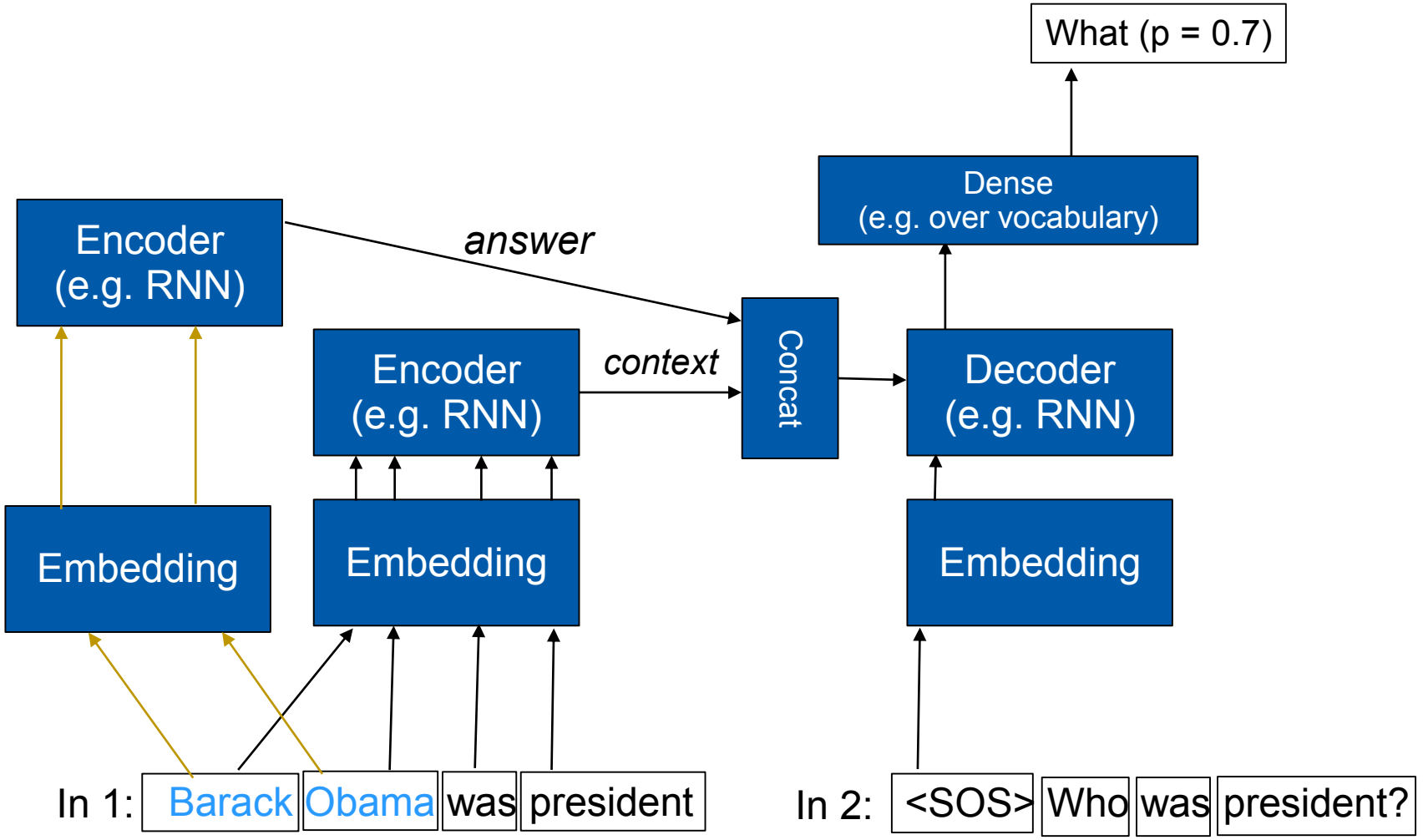


Statistical Question Generation(2)
Training Example

learn $P(w_i = w | w_{i-1}, \dots, w_0, context, answer)$

Gold: Who was president?

≠



Statistical Question Generation: Inference

TECHNISCHE UNIVERSITÄT DARMSTADT

During inference

Use learnt $P(w_i = w | w_{i-1}, \dots, w_0, context, answer)$

data

~~gold questions~~ answers context sentences

Statistical Question Generation (3)
Inference Example

TECHNISCHE UNIVERSITÄT DARMSTADT

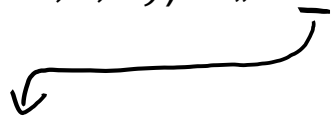
Inference example

Context *c*:

"The end-to-end delay should be as small as possible, in particular for applications working in dialog, similar to the conventional phone"

Answer *a*:
end-to-end delay

$argmax(P(w_1 = w|\, SOS, c, a))$ = „What"



$argmax(P(w_2 = w|\, What, SOS, c, a))$ = „is"

…

$argmax(P(w_8 = w|\, for, \ldots, What, SOS, c, a))$ = „?"

Possible result:
"What is the end-to-end delay good for ?"

Example generated questions

TECHNISCHE UNIVERSITÄT DARMSTADT

Sentence:

While most people were warming up their cars , Trevor , my husband , had to get up early to ride his bike four kilometers away from home to work.

Question:

Where were most people warming up ?

Sentence:

Here 's some advice about how to choose the right diet to keep your eyes healthy.

Question:

What is the name of the advice that you get from your doctor ?

EVALUATION OF AQG SYSTEMS

NLG Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

	intrinsic	extrinsic
automatic	▪ N-gram overlap	????
manual	▪ human ratings of questions	▪ Human task performance

Automatic Evaluation: BLEU Metric [9]

TECHNISCHE UNIVERSITÄT DARMSTADT

Reference 1	Who	let	the	happy	dogs	out	?
Reference 2	Who	released	the	happy	hounds	?	
Candidate	Who	released	the	forks	out	out	?

Automatic Evaluation: BLEU Metric [9]

TECHNISCHE UNIVERSITÄT DARMSTADT

Reference 1	Who	let	the	happy	dogs	out	?
Reference 2	Who	released	the	happy	hounds	?	
Candidate	Who	released	the	forks	out	out	?

Would you say the candidate is a good generated question based on the references?

Example: BLEU Metric [9]

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

Reference 1	Who	let	the	happy	dogs	out	?
Reference 2	Who	released	the	happy	hounds	?	
Candidate	Who	released	the	forks	out	out	?

Example: BLEU Metric [9]

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

Reference 1	Who	let	the	happy	dogs	out	?
Reference 2	Who	released	the	happy	hounds	?	
Candidate	Who	released	the	forks	out	out	?

clipped clipped

$$\frac{1+1+1+0+1+1}{1+1+1+1+1+1+1} = \frac{5}{7}$$

Candidate has 7 1-grams

Automatic Evaluation: BLEU Metric [9]

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} Count_{clip}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} Count(n\text{-}gram')}.$$

For corpus level:
Adjust for short answers
And weight n-grams

Reference 1	Who	let	the	happy	dogs	out	?
Reference 2	Who	released	the	happy	hounds	?	
Candidate	Who	released	the	forks	out	out	?

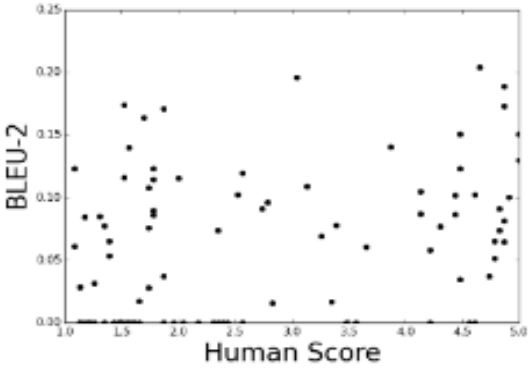
Problems of Automatic Metrics

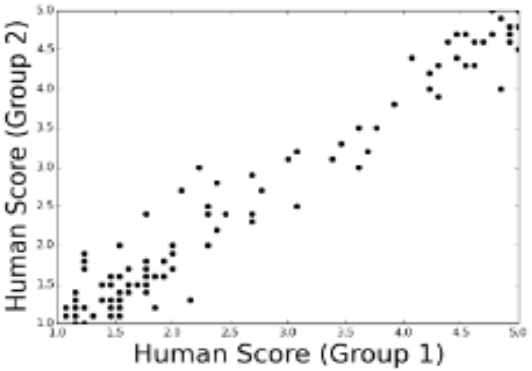
TECHNISCHE UNIVERSITÄT DARMSTADT

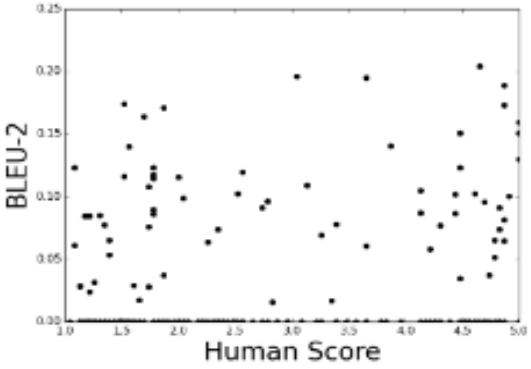
- **Need gold data**
- **Measure only structure (a little)**
- **Do not measure the important scales:**
  - Answerability
  - Relevance
  - Pedagogical value
  - …

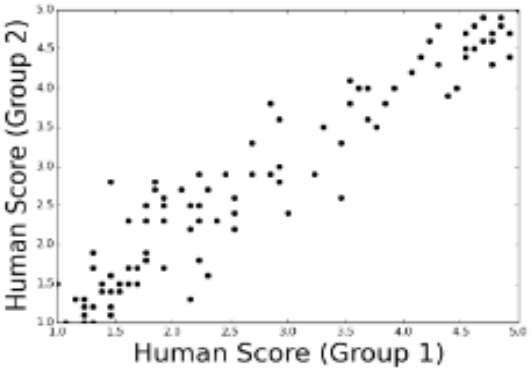
Problems of Automatic Metrics

- **Need gold data**
- **Measure only structure (a little)**
- **Do not measure the important scales:**
  - Answerability
  - Relevance
  - Pedagogical value
  - …









[11]

Manual Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

Intrinsic

Quality of the generated questions

- Grammaticality
- Relevance
- Answerability

Manual Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

Intrinsic

Quality of the generated questions

- Grammaticality
- Relevance
- Answerability

← How would a corresponding study look like? What are challenges?

Manual Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

Intrinsic

Quality of the generated questions

- Grammaticality
- Relevance
- Answerability

Extrinsic

Usefulness of the questions for the upstream task

- Learning outcome
- Motivation
- Time on task

WRAP UP

We Have Covered

TECHNISCHE UNIVERSITÄT DARMSTADT

Questions in Education

- Ample evidence that they help
- Different approaches for different purposes

Automatic Question Generation

- Involves multiple steps
- LexRank for question-worthy sentence selection
- Seq2Seq models for output generation

Evaluation Methods

- Automatic evaluation
- Automatic BLEU scores
- Manual evaluation

About the Exercise

TECHNISCHE UNIVERSITÄT DARMSTADT

Task1 (3p)

- Further understanding of LexRank

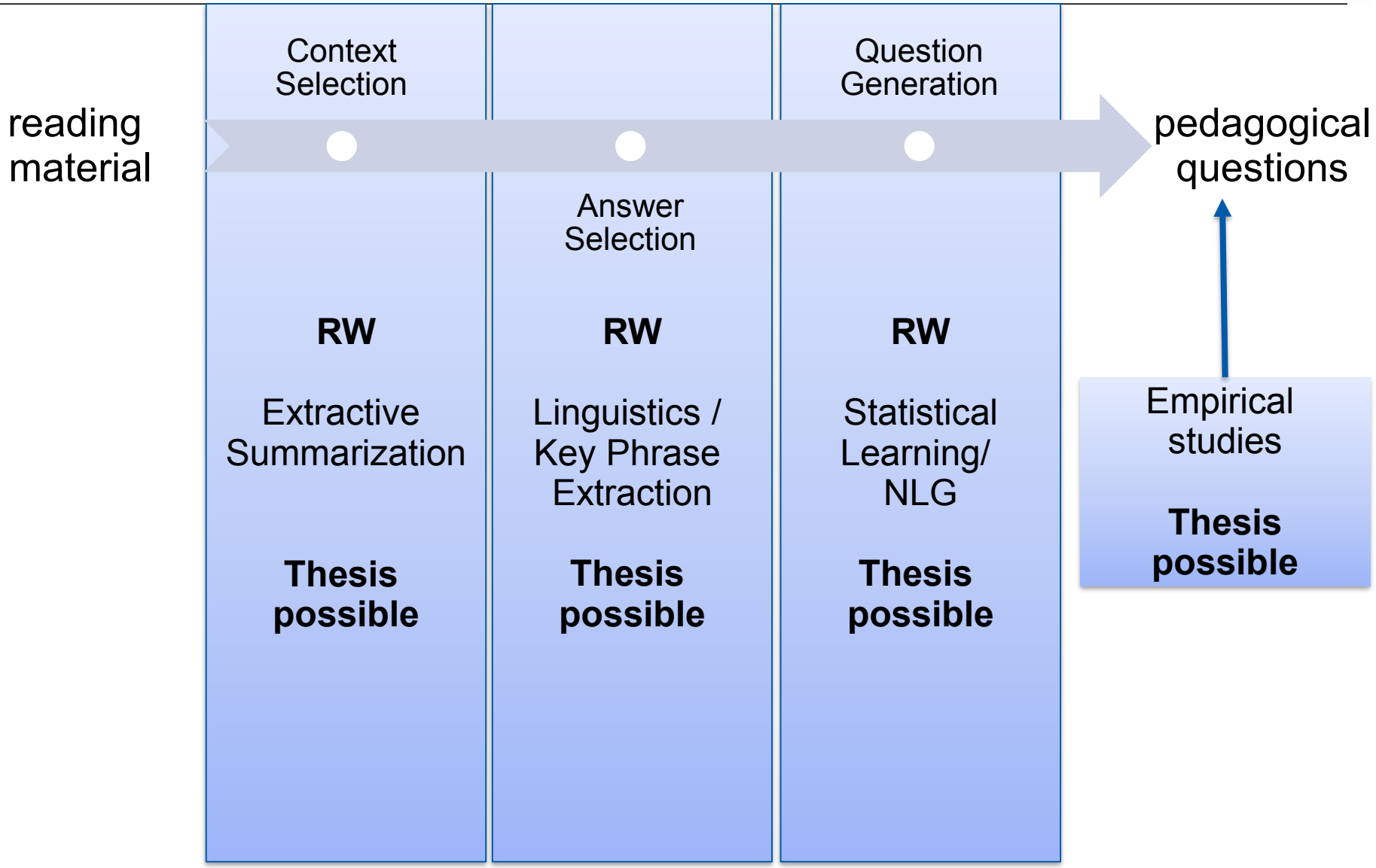
Task 2 (3p)

- Further understanding of Neural Networks for NLG

Task 3 (1p)

- Calculating BLEU

Want to dive deeper?



Endslide

Department of Electrical Engineering
and Information Technology
Multimedia Communications Lab - KOM

TECHNISCHE UNIVERSITÄT DARMSTADT

Tim Steuer, M.Sc.

Tim.Steuer@KOM.tu-darmstadt.de
Rundeturmstr. 10
64283 Darmstadt/Germany
www.kom.tu-darmstadt.de

Phone +49 6151 16-20463
Fax +49 6151 16-29109

References

1. Le, N. T., Kojiri, T., & Pinkwart, N. (2014). Automatic question generation for educational applications–the state of art. In *Advanced Computational Methods for Knowledge Engineering* (pp. 325-338). Springer, Cham.
2. Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In *Psychology of learning and motivation* (Vol. 9, pp. 89-132). Academic Press.
3. Ausubel, D. P. (1962). A subsumption theory of meaningful verbal learning and retention. *The Journal of General Psychology*, *66*(2), 213-224.
4. Bloom, B. S. (1973). Recent developments in mastery learning. *Educational Psychologist*, *10*(2), 53-57.
5. VanLehn, K., Lnch, C., Schulze, K., Shapiro, J. A., Shelby, R., Taylor, L., ... & Wintersgill, M. (2005). *The Andes physics tutoring system: Five years of evaluations*. Naval Academy Annapolis MD.
6. VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading?. *Cognitive science*, *31*(1), 3-62.
7. Chen, G., Yang, J., & Gasevic, D. (2019, June). A Comparative Study on Question-Worthy Sentence Selection Strategies for Educational Question Generation. In *International Conference on Artificial Intelligence in Education* (pp. 59-70). Springer, Cham.
8. Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, *22*, 457-479.
9. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.
10. Ritter, S., Yudelson, M., Fancsali, S. E., & Berman, S. R. (2016, April). How mastery learning works at scale. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 71-79). ACM.
11. Liu, C. W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., & Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.