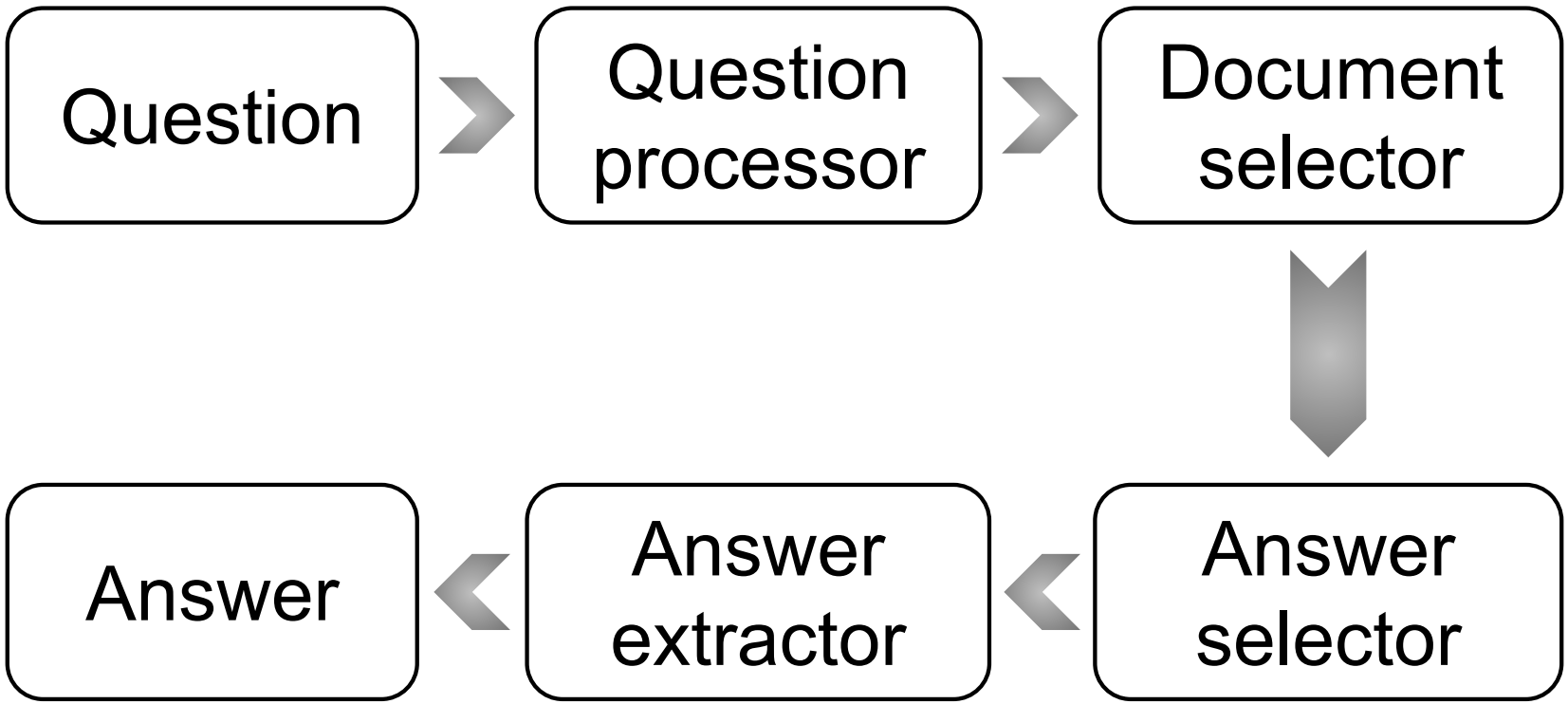
Question Answering

Marek Medveď, Aleš Horák

Centrum ZPJ, FI MU, Brno

January 17, 2019

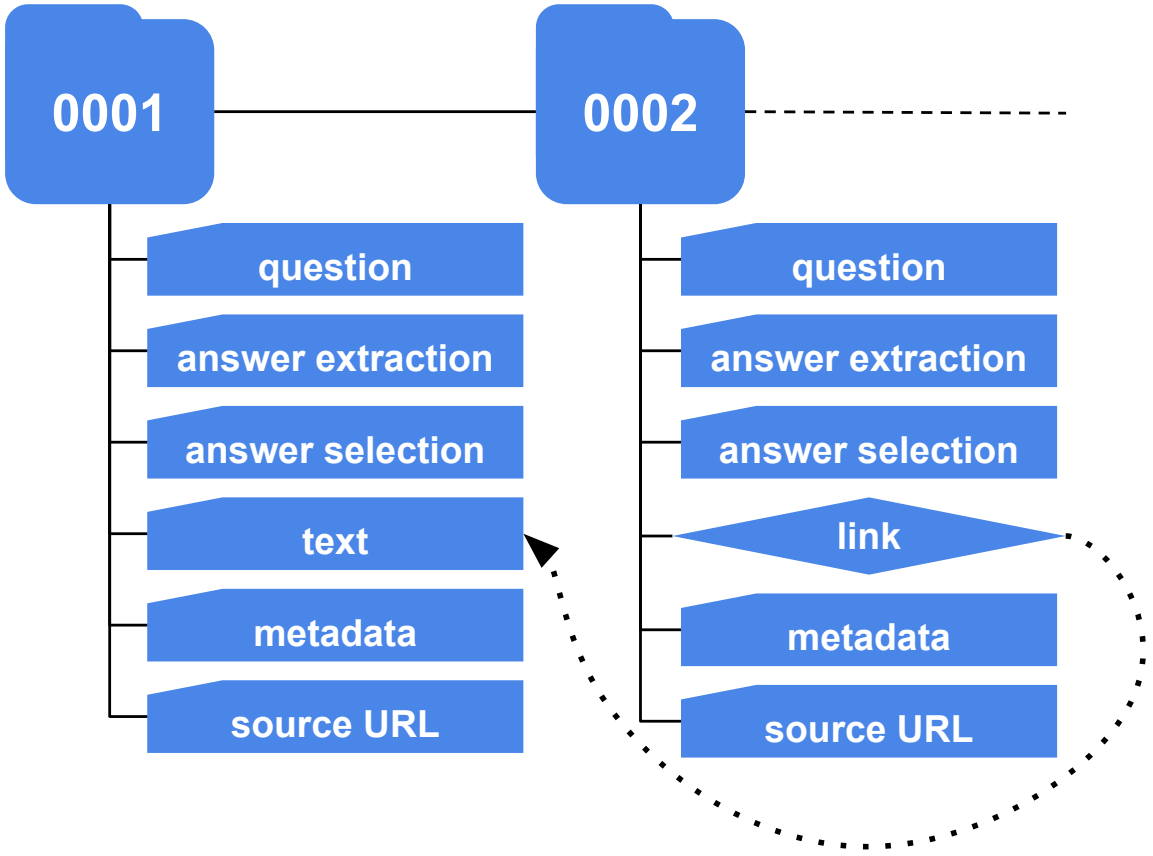
Automatic Question Answering system (AQA)



SQAD v2.1

SQAD v2.1

Czech Simple Question Answering Dataset
8,566 questions with answers and metadata



all texts are supplemented with Part-of-Speech tags and lemmata

SQAD v2.1 example

Example

`01question.txt`:

Ve kterém roce byla založena hudební skupina Polemic?

`02answer.txt`:

1989

`06answer.selection.txt`:

Polemic je osmičlenná slovenská hudební skupina založená v roce 1989 v Bratislavě.

`03text.extended.txt`:

Polemic je osmičlenná slovenská hudební skupina založená v roce 1989 v Bratislavě. Patří mezi přední slovenské skupiny hrající styly ska a reggae. Velmi oblíbení jsou i v České republice. ...

`04url.txt`:

`http://cs.wikipedia.org/wiki/Polemic`

`05metadata.txt`:

```
<q_type>DATETIME</q_type>
<a_type>DATETIME</a_type>
```

Question types

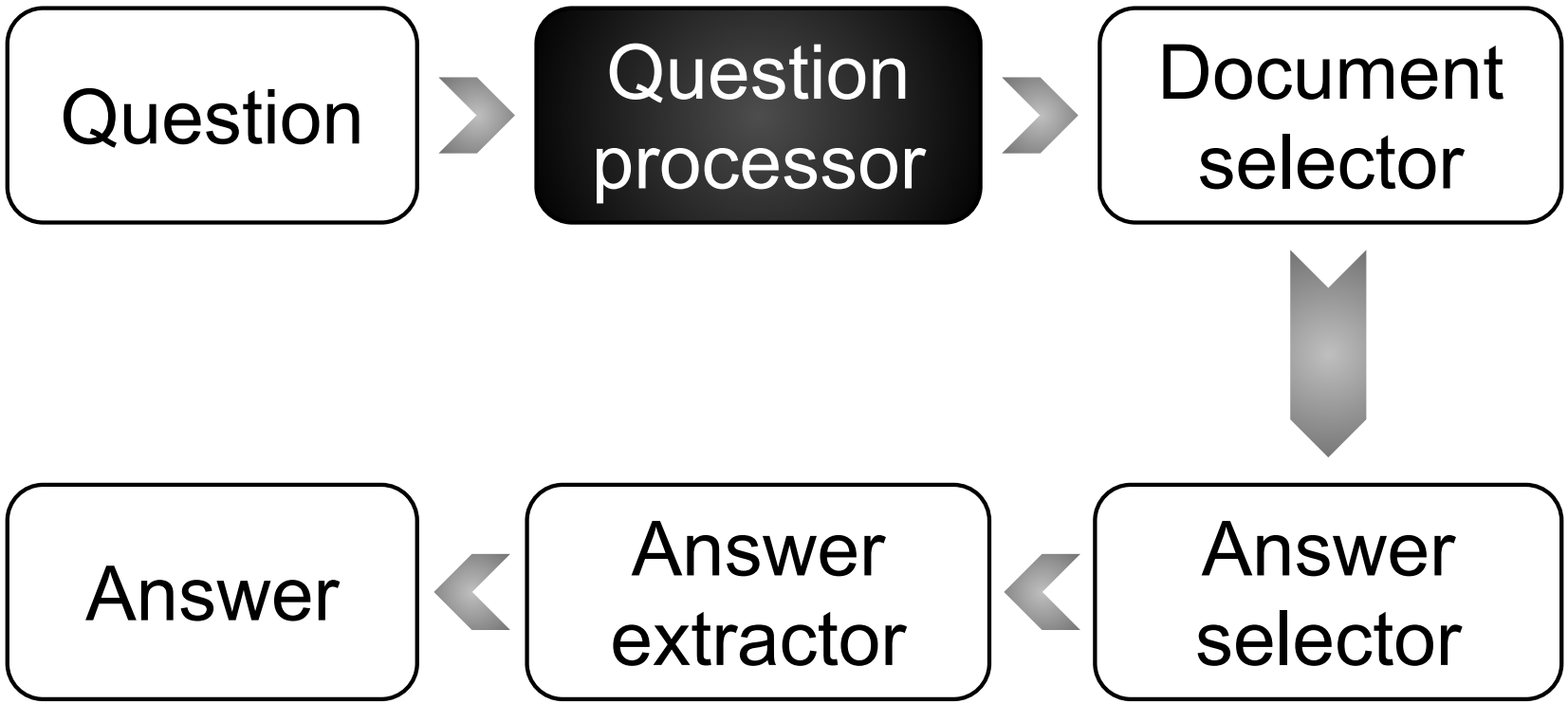
Database	SQAD v2.0	SQAD v2.1
PERSON	940	1,023
ENTITY	1,436	1,745
ADJ_PHRASE	253	233
DATE/TIME	1,848	1,851
LOCATION	1,436	1,524
NUMERIC	900	913
ABBREVIATION	-	81
CLAUSE	774	241
VERB_PHRASE	944	940
OTHER	31	15

Answer types

Database	SQAD v2.0	SQAD v2.1
PERSON	943	1,050
DENOTATION	-	102
ENTITY	811	1,085
OTHER	1,480	819
ORGANIZATION	199	216
DATE/TIME	1,847	1,845
LOCATION	1,442	1,511
NUMERIC	904	918
ABBREVIATION	-	82
YES/NO	940	938

Q/A-type Detection

AQA system: module position



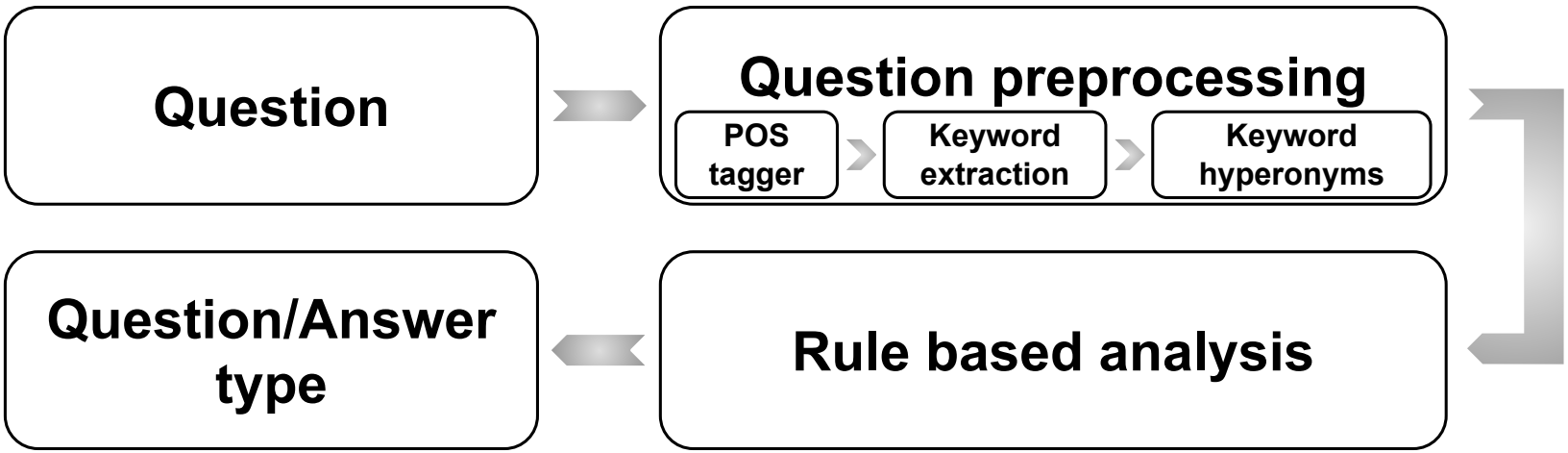
Original implementation

Example

(AGENT, LANGUAGE, PLACE, WORK): [ktery, jaky, prepktery, ...]
(AGENT, PLACE, NEUTRAL): [kdo, koho, komu, kym, ...]
(AGENT, LANGUAGE, WORK, NEUTRAL): [cim]
(AGENT, PLACE): [kde, kam, sestranyceho, zakoho]
PLACE: [kudy]
T: [kdy, kolik, zkolika, dokdy, dokolika, ...]
NEUTRAL: [proc, zaceho, naproticemu]
DEATH/BIRTH: [kdebirth, kdybirth, kdydeath, kdedeath]

- recognition through SET (Matej Pavla 2014)

Question type analysis (Rule based)



Keyword extraction (Rule based)

- First noun after the relative pronoun *"který"* (which) or *"jaký"* (what), **NOT** part of a relative sentence.
- First noun after the first verb
- First following noun after: *"název"* (title), *"pojem"* (concept), *"termín"* (term), *"typ"* (type), *"část"* (part), or *"větev"* (branch)

Keyword hypernyms (Rule based)

- obtained by means of the Czech Wordnet API
  1. API is queried to find **all possible senses**
  2. API is queried to create a list of hypernyms for **three most common word senses**

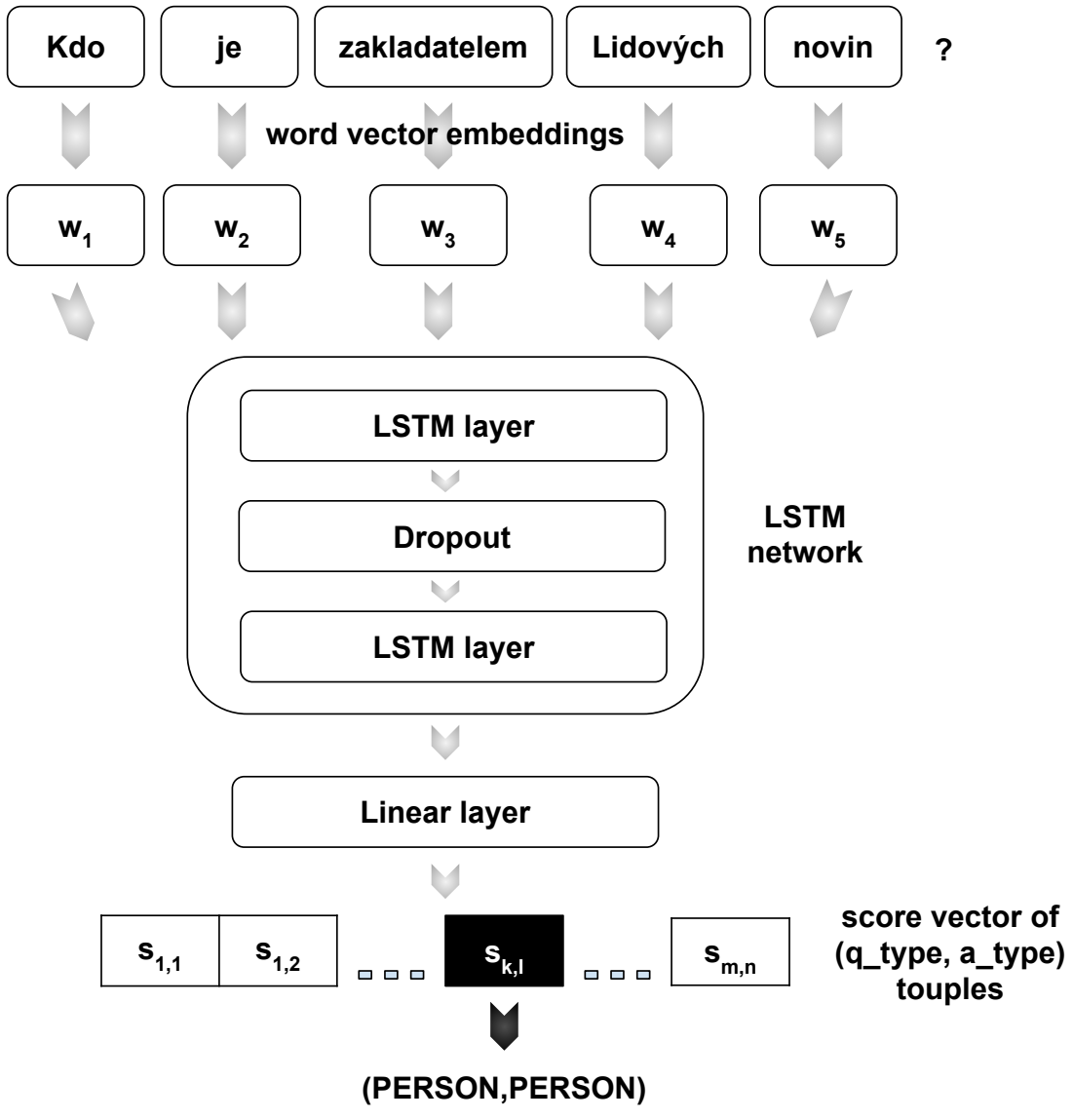
Features recognised by rules (Rule based)

- keyword hypernym match:
  Example: `"<word>" in keyword.hypernym`
- important word recognition:
  Example: `"<word>" == words.lemma_at_index(0)`
  $\rightarrow$ the first word in the sentence is the specified word
- question structure match:
  Example: `"k2" in words.tag_at_index(1)`
  $\rightarrow$ the second word in the sentence is an adjective

Example

question:	Jak se jmenovala první manželka Miloše Formana? (What was the name of the first wife of Miloš Forman?)
keyword:	manželka (wife)
hypernyms:	[manželka, jednotlivec, osoba, bytost, organismus] (wife, individual, person, being, organism)
rule:	`(PERSON; PERSON)` → `"osoba" in keyword.hypernym`

Question type analysis (Machine learning based)



SQAD subsets in evaluation

	training	evaluation	testing
Rule based system	4,279	-	4,287
LSTM network	7,011	735	820

Evaluation: rule based

Rule based	precision	recall	F1
question type	88.77%	87.79%	88.28%
answer type	**85.05%**	**84.52%**	**84.78%**
both types	82.43%	82.93%	82.68%

Evaluation: machine learning based

Model training setup: 40 epochs, batch size of 64, dropout rate of 0.5 and learning rate of 0.001

ML based	precision	recall	F1
question type	91.59%	90.73%	91.16%
answer type	**89.76%**	**89.14%**	**89.45%**
both types	86.15%	87.07%	86.61%

Rule based			
question type	88.77%	87.79%	88.28%
answer type	**85.05%**	**84.52%**	**84.78%**
both types	82.43%	82.93%	82.68%

Question type confusion matrix: rule based

	expected									
pred	AB	APHR	CLS	D/T	ENT	LOC	NUM	OTH	PER	VPHR
AB	37	1	1	0	19	3	1	0	0	0
APHR	1	52	4	0	49	6	6	0	4	0
CLS	1	0	35	0	14	4	0	0	5	0
D/T	0	0	1	916	16	0	2	0	1	1
ENT	0	44	71	3	685	41	13	2	40	8
LOC	0	6	1	0	22	695	3	0	3	1
NUM	1	4	1	4	8	0	422	0	0	0
OTH	0	1	3	2	25	7	7	5	3	6
PER	0	8	3	0	33	6	2	0	455	0
VPHR	0	0	0	0	0	0	0	0	0	454

Answer type confusion matrix: rule based

	expected									
pred	AB	D/T	ENT	LOC	NUM	ORG	OTH	PER	DNT	Y/N
AB	37	0	9	3	1	1	9	2	0	0
D/T	0	915	7	0	2	1	8	1	2	1
ENT	0	2	405	32	14	19	191	40	10	5
LOC	0	0	7	693	3	9	15	3	0	1
NUM	1	3	3	0	423	0	9	0	1	0
ORG	1	0	30	5	0	61	24	6	0	0
OTH	2	2	46	16	14	10	138	19	3	7
PER	0	0	12	7	2	13	18	452	3	0
DNT	0	0	1	1	1	1	3	0	38	0
Y/N	0	0	0	0	0	0	0	0	0	454

Question type confusion matrix: machine learning based

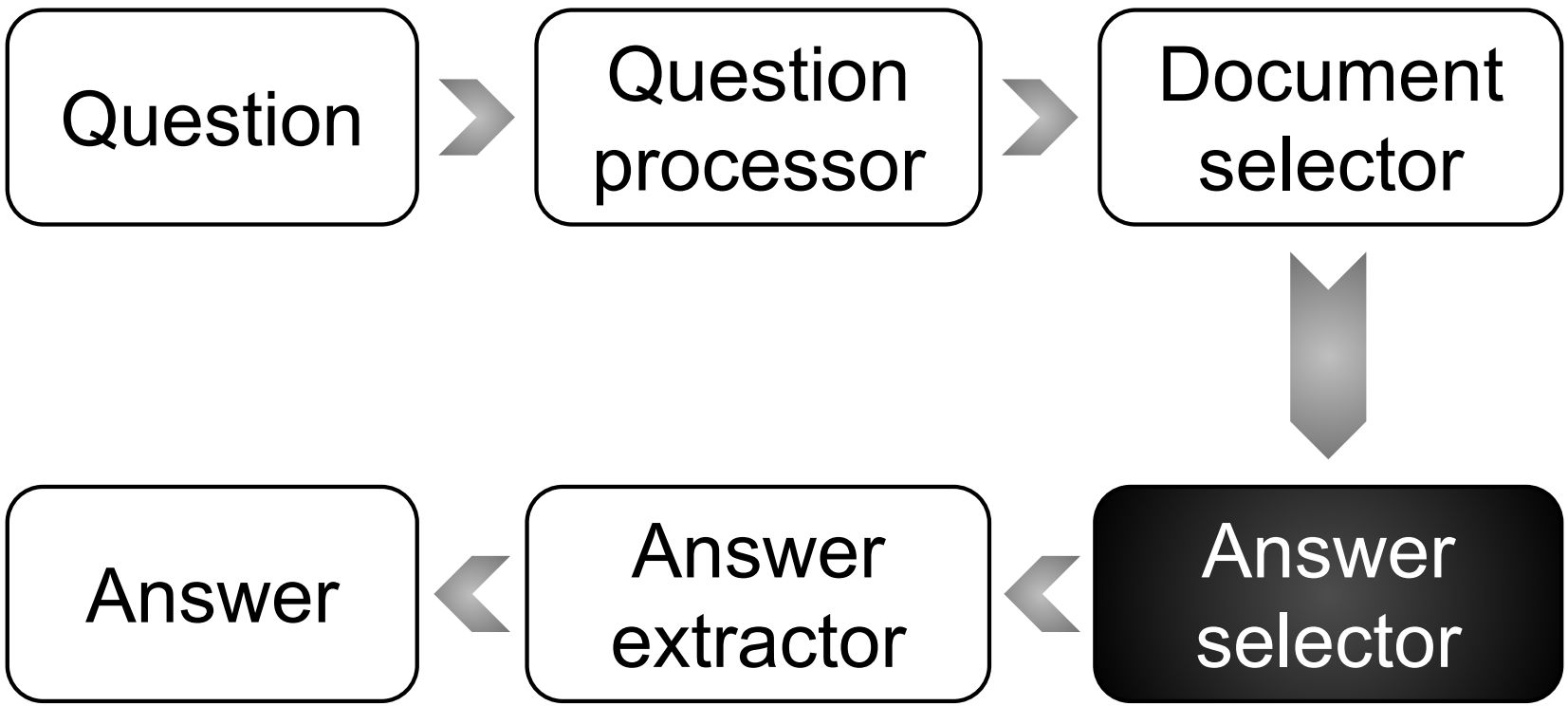
	expected									
pred	AB	APHR	CLS	D/T	ENT	LOC	NUM	OTH	PER	VPHR
AB	7	0	0	0	2	0	0	0	0	0
APHR	0	12	0	0	9	2	0	0	1	0
CLS	0	0	9	0	12	0	0	0	3	0
D/T	0	0	0	175	0	0	0	0	0	0
ENT	0	5	6	0	129	3	1	0	9	1
LOC	0	1	0	0	7	141	0	0	1	0
NUM	1	1	0	1	0	0	87	0	0	0
OTH	0	0	0	0	1	0	0	0	0	1
PER	0	0	0	0	6	0	0	0	95	0
VPHR.	0	0	0	0	1	0	0	1	0	89

Answer type confusion matrix: machine learning based

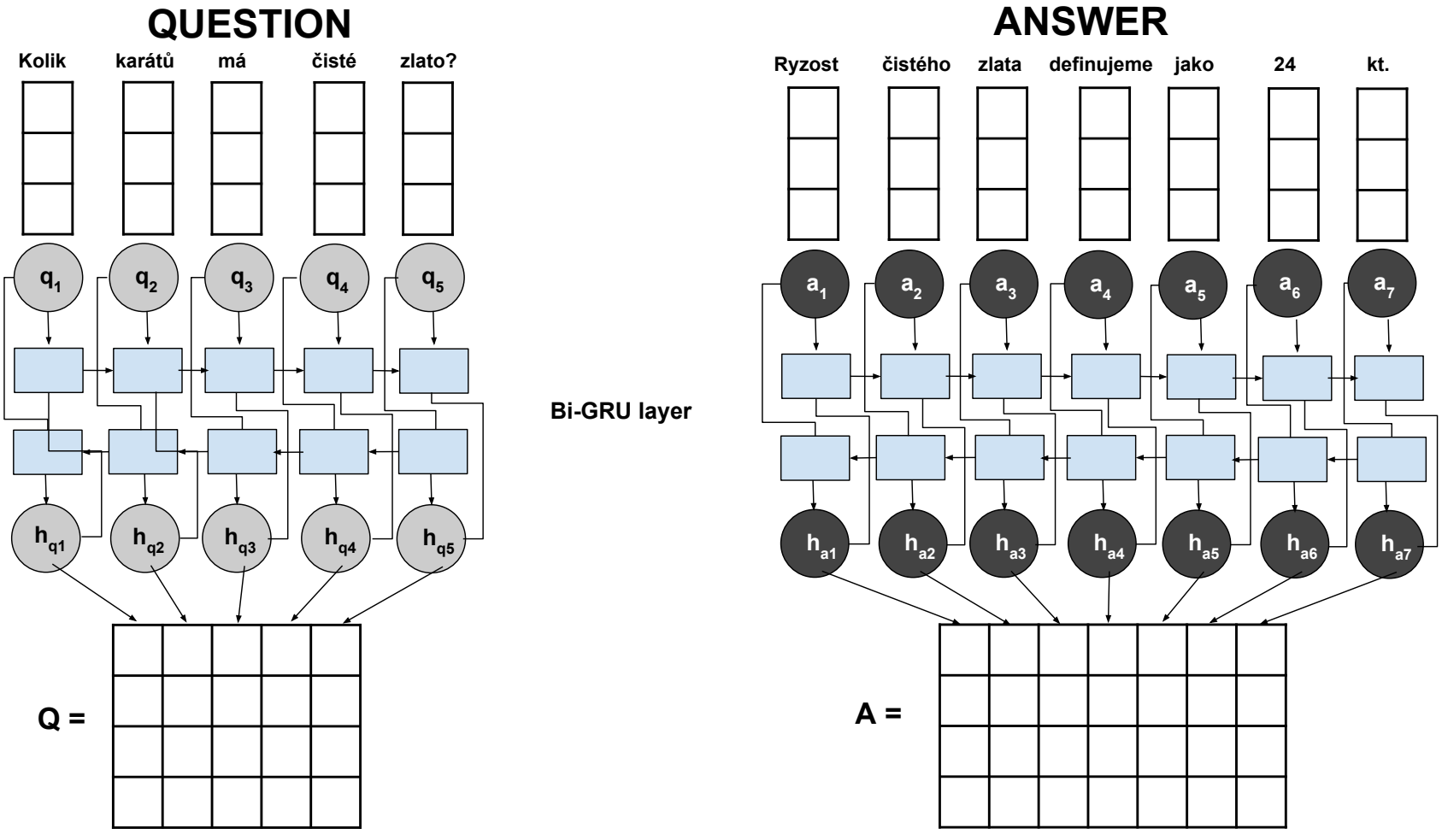
	expected									
pred	AB	D/T	ENT	LOC	NUM	ORG	OTH	PER	DNT	Y/N
AB	7	0	0	0	0	1	1	0	0	0
D/T	0	175	0	0	0	0	0	0	0	0
ENT	0	1	72	4	0	2	13	3	1	0
LOC	0	0	2	140	0	2	2	1	0	0
NUM	1	1	0	0	87	0	1	0	0	0
ORG	0	0	1	0	0	12	1	3	0	0
OTH	0	3	20	2	1	0	44	7	2	2
PER	0	0	3	0	0	2	2	96	0	0
DNT.	0	0	2	0	0	0	1	0	9	0
Y/N	0	0	0	0	0	0	1	0	0	89

Answer selection

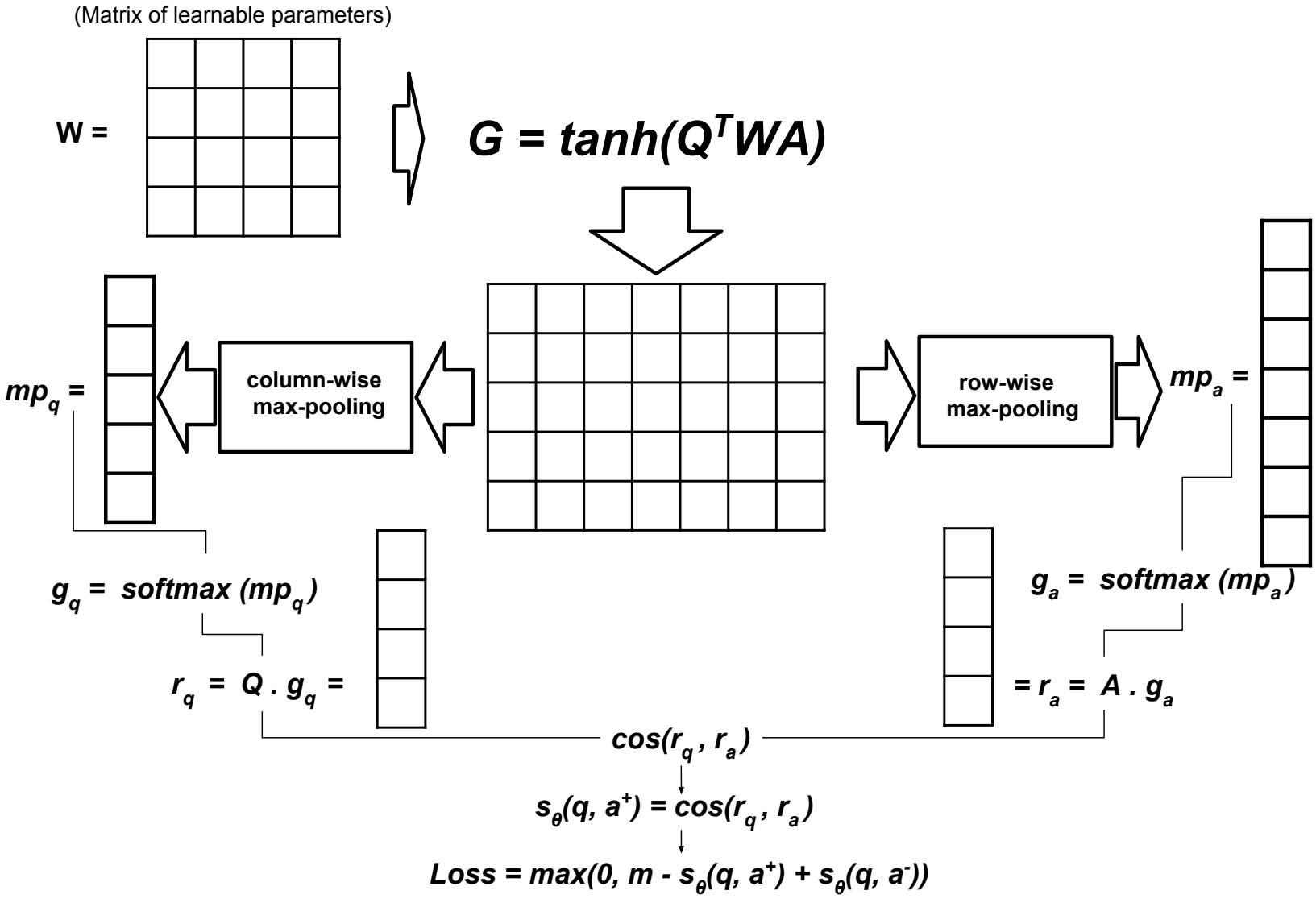
AQA system


Question
Question processor
Document selector
Answer selector
Answer extractor
Answer

Answer selection model



Answer selection model

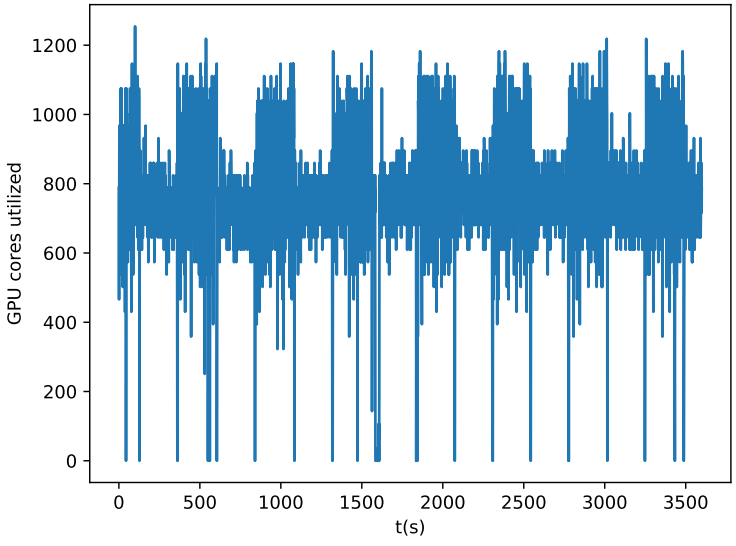


Training

- 25 epochs
- 50 randomly chosen answers for each question
- $Loss = max\{0, m - s_\theta(q, a^+) + s_\theta(q, a^-)\}$
- m – constant margin (0.2)
- s_θ – the cosine similarity as computed by the network with parameters θ, q is the input question and a^+/a^- are the positive/negative answers.

GPU Utilization

a) one model



b) two models at the same time

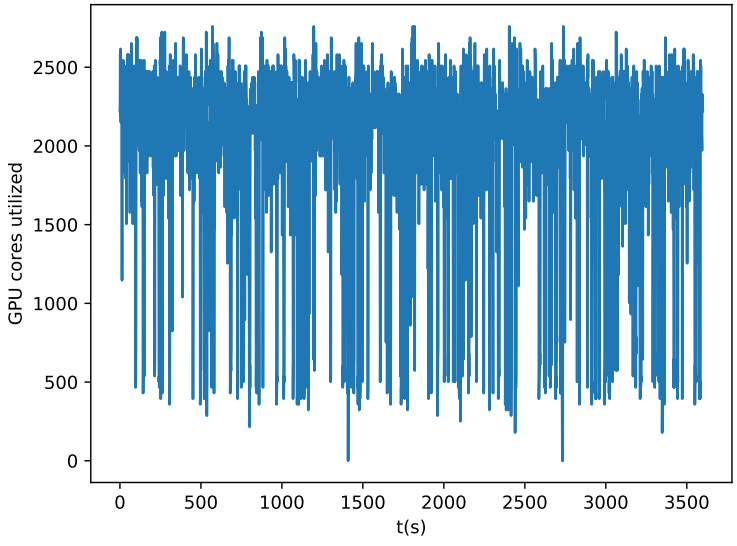
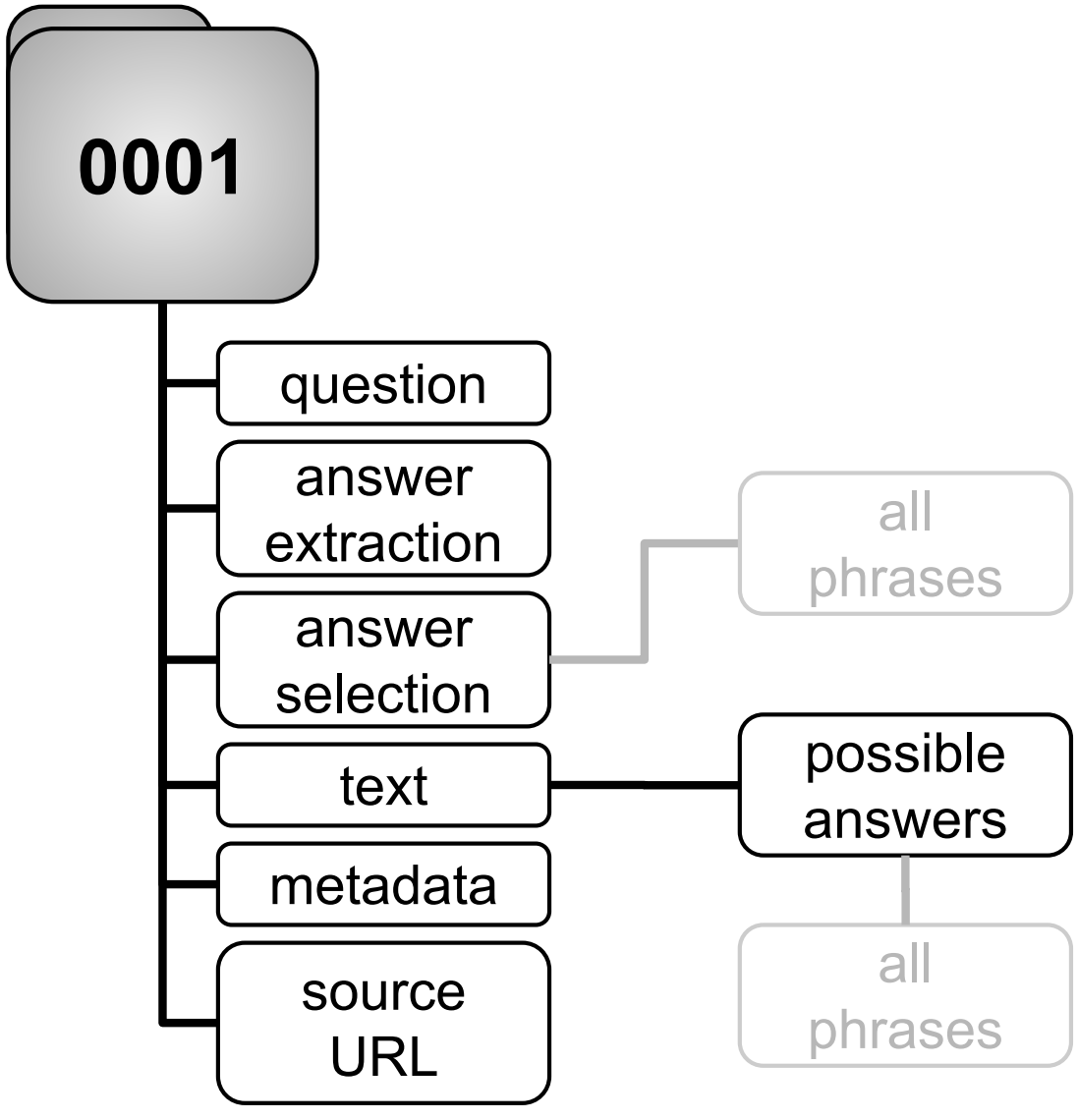


Figure: GPU core utilization for one hour of training. Maximum number of working GPU cores is 3584.

- CPU running time for epoch – approx. 3 hours
- GPU running time for epoch – 380 seconds, 28 $*$ improvement

SQAD adjustments



Model evaluation

Table: The results for combinations of hyperparameter values

Output size	Dropout	Learning rate λ	Test set (in %)
260	0.2	0.05	61.89
		0.1	65.47
		0.2	**68.29**

Improved answer selection results

	old answer selection	current answer selection
1	56.53 %	**66.03 %**
2	8.21 %	4.27 %
3	4.00 %	1.33 %
4	2.42 %	0.73 %
5	1.97 %	0.27 %
6	1.51 %	0.15 %
7	1.51 %	0.06 %
8	1.73 %	0.15 %
9	0.85 %	0.12 %
10	1.03 %	0.15 %
>10	20.24 %	26.75 %

Answer selection results for particular question-answer types

q-a type	total	first pos.	accuracy
LOCATION - OTHER	2	2	100.00%
ADJ_PHRASE - PERSON	2	2	100.00%
LOCATION - ORGANIZATION	1	1	100.00%
DATETIME - NUMERIC	1	1	100.00%
ABBREVIATION - ABBREVIATION	32	28	87.50%
LOCATION - LOCATION	600	464	77.33%
ENTITY - ORGANIZATION	81	60	74.07%
DATETIME - DATETIME	737	538	73.00%
VERB_PHRASE - YES_NO	374	254	67.91%
ENTITY - OTHER	151	101	66.89%
ADJ_PHRASE - ENTITY	3	2	66.67%
PERSON - PERSON	406	265	65.27%

Answer selection results for particular question-answer types

q-a type	total	first pos.	accuracy
ADJ_PHRASE - OTHER	86	56	65.12%
ENTITY - TEACHING	40	26	65.00%
NUMERIC - NUMERIC	364	228	62.64%
ENTITY - PERSON	8	5	62.50%
ENTITY - ENTITY	412	249	60.44%
LOCATION - ENTITY	5	3	60.00%
CLAUSE - OTHER	82	38	46.34%
OTHER - OTHER	5	2	40.00%
CLAUSE - ENTITY	11	2	18.18%
PERSON - ORGANIZATION	1	0	0.00%
ENTITY - LOCATION	2	0	0.00%
CLAUSE - ORGANIZATION	2	0	0.00%
sum	3408	2327	68.29%

Example (Correctly answered question)

Question: Jak nazval Kolumbus obyvatele ním objevené země?
Chosen answer: Kolumbus nevěděl, že objevil pro Evropu nový kontinent a obyvatele objevených zemí nazval Indios (španělský výraz pro Indy).
- **SCORE = 0.71075**

Candidate answers that ended up on 2nd and 3rd place:
Kolumbus nebyl prvním Evropanem, který navštívil Ameriku – o pět století dříve jejích břehů dosáhla norská expedice vedená Leifem Erikssonem, která založila kolonii na dnešním Newfoundlandu.
- **SCORE = 0.60405**

V roce 1538 nazval Gerhard Mercator Amerikou celý kontinent.
- **SCORE = 0.53724**

Example (Incorrectly answered question)

Question: Má pivo příznivé účinky?
Chosen incorrect answer: Současně výrazně podporuje chuť k jídlu, což může vést při nestřídmé konzumaci pokrmů k nárůstu tělesné hmotnosti.
- **SCORE = 0.66412**

Right answer (that ended up in 5th position):
Uvádí se, že konzumace piva má příznivé účinky na dobrou náladu, podporu krevního oběhu, snížení rizika srdečních příhod a působí proti vysokému krevnímu tlaku.
- **SCORE = 0.64472**

Example (Incorrectly answered question)

Candidate answers that ended up between chosen answer and correct answer:

Ovšem i střídmé pití však pravděpodobně škodí, protože studie jsou často ovlivněny vlivy systematických chyb, kdy do skupiny abstinentů například spadají lidé, kteří nepijí ze zdravotních důvodů.
- **SCORE= 0.65227**

Barvení kulérem (karamelem) nebo pražením sladu je sice pro pivo charakteristické, avšak antioxidační hodnotu piva nezvyšuje.
- **SCORE= 0.64678**

Příznivé účinky piva na lidský organismus se mohou projevit při jeho střídmé konzumaci, kdy nepřevažují negativní účinky alkoholu.
- **SCORE= 0.64673**

Future work

Future work

- Improve document selection module
- Connect new modules into current pipeline
- Experiment with answer selection module setup
- Experiment with (automatic) TIL constructions of questions and answers to improve phrase matching

Karel Pepper