Encoder-based Jazykové Modely v Seznamu

Jakub Náplava

Náš život s velkými i malými jazykovými modely

Seznam.cz


Seznam.cz


Televize Seznam


stream


sreality.cz


SPORT.CZ


EXPRESFM


email


sdovolená.cz


sblog


sklik.cz


sauto.cz


smoto.cz


Novinky.cz


mapy.cz


Kupi.cz


PROŽENY


SUPER.CZ


Zboží.cz


Firmy.cz


Seznam Zprávy


CZECH
PROPERTIES


sbazar.cz


HRY.cz

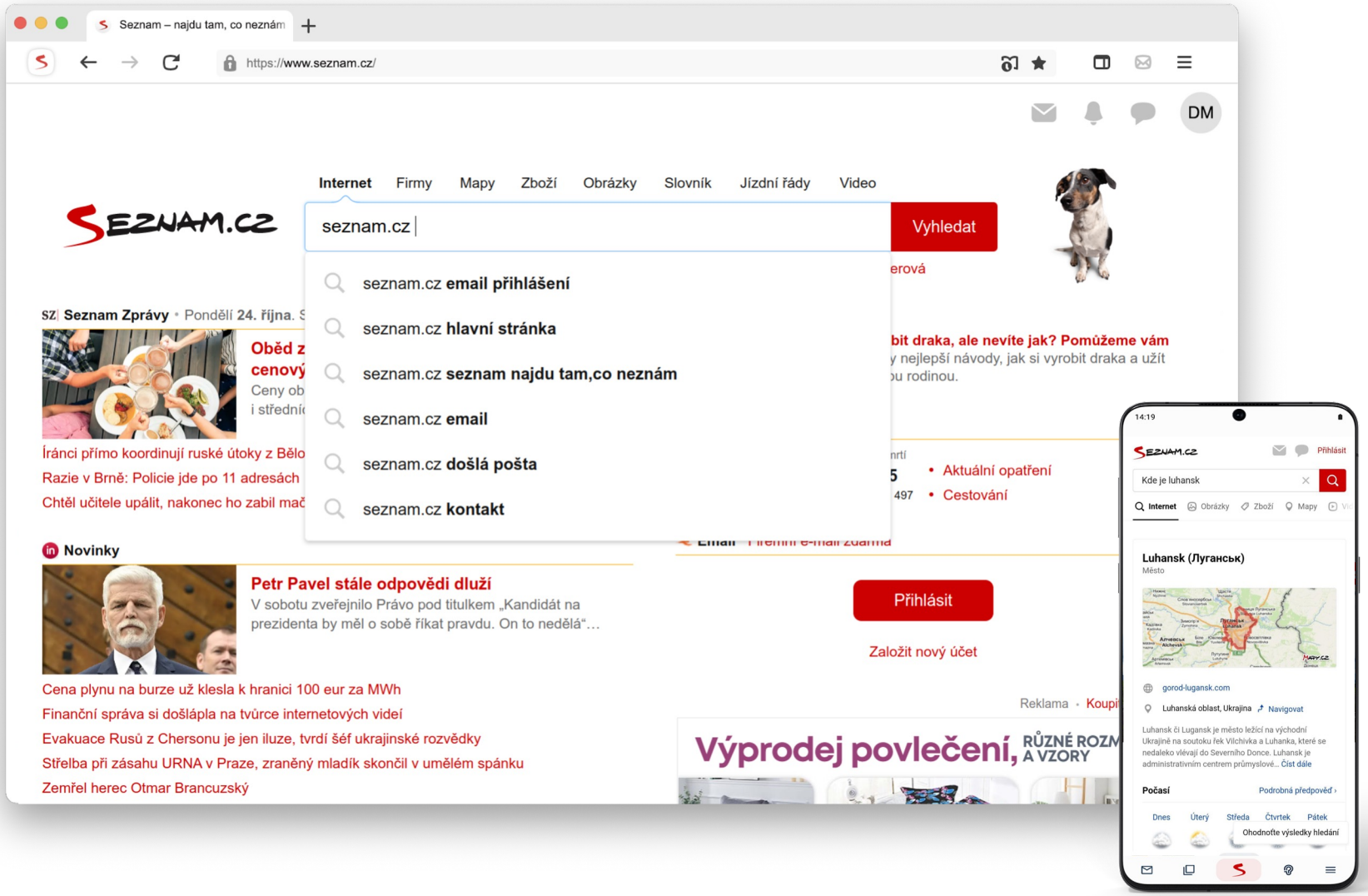

classic
praha


Horoskopy.cz


GARÁŽ.CZ


POČASÍ.CZ


LIDÉ


Seznam – najdu tam, co neznám
https://www.seznam.cz/
DM
Internet
Firmy
Mapy
Zboží
Obrázky
Slovník
Jízdní řády
Video
SEZNAM.CZ
seznam.cz
Vyhledat
seznam.cz email přihlášení
seznam.cz hlavní stránka
seznam.cz seznam najdu tam,co neznám
seznam.cz email
seznam.cz došlá pošta
seznam.cz kontakt
SZ Seznam Zprávy • Pondělí 24. října.
Íránci přímo koordinují ruské útoky z Bělo
Razie v Brně: Policie jde po 11 adresách
Chtěl učitele upálit, nakonec ho zabil mač
Novinky
Petr Pavel stále odpovědi dluží
V sobotu zveřejnilo Právo pod titulkem „Kandidát na prezidenta by měl o sobě říkat pravdu. On to nedělá“…
Cena plynu na burze už klesla k hranici 100 eur za MWh
Finanční správa si došlápla na tvůrce internetových videí
Evakuace Rusů z Chersonu je jen iluze, tvrdí šéf ukrajinské rozvědky
Střelba při zásahu URNA v Praze, zraněný mladík skončil v umělém spánku
Zemřel herec Otmar Brancuzský
Aktuální opatření
Cestování
Přihlásit
Založit nový účet
Reklama
Výprodej povlečení,
14:19
Kde je luhansk
Internet
Obrázky
Zboží
Mapy
Luhansk (Луганськ)
Město
gorod-lugansk.com
Luhanská oblast, Ukrajina
Navigovat
Počasí
Podrobná předpověď
Dnes
Úterý
Středa
Čtvrtek
Pátek
Ohodnoťte výsledky hledání

Contents

- Encoder Language Models in Seznam.cz
- Czech Semantic Embedding Models
- Future Plans

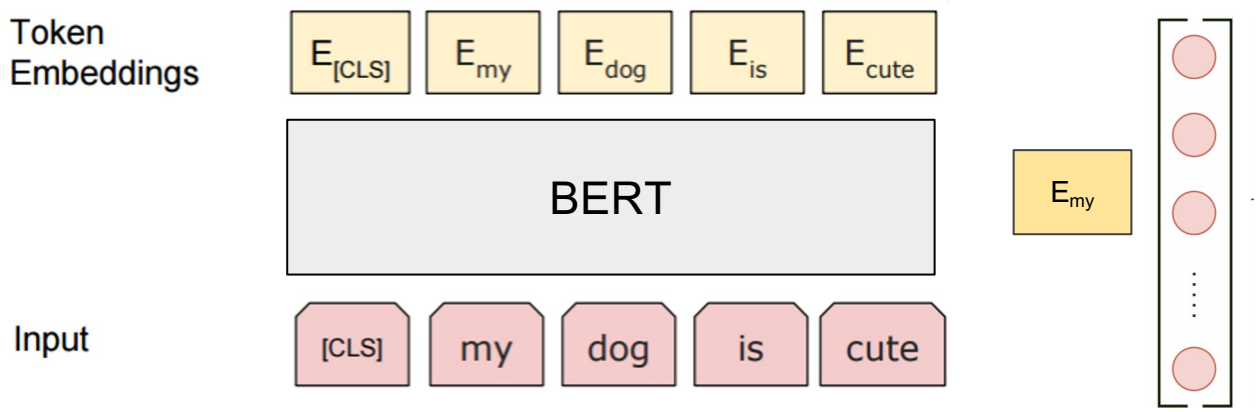
Encoder based Language Models in Seznam.cz

What are Encoder Language Models?

Neural models that model text and can learn complex text tasks (that were impossible to do before).

BERT: Pre-training of Deep Bidirectional Transformers

for Language Understanding (2018)



Seznam & Text

Vyhledávání

Seznam.cz

Inzerce

Sklik.cz

Oborové služby

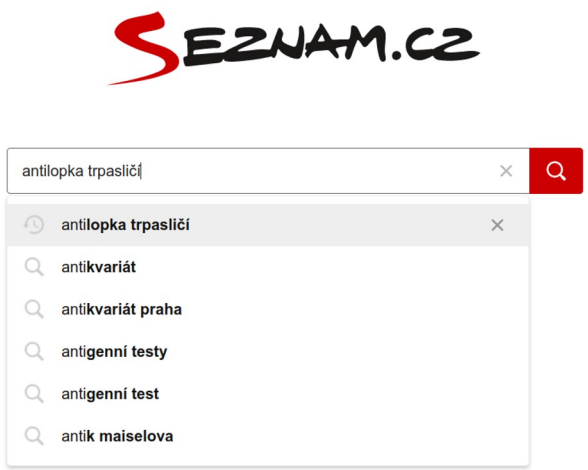
Zboží.cz | Sreality.cz | Sauto.cz | Sbazar.cz | Smoto.cz | Sdovolená.cz | volná místa

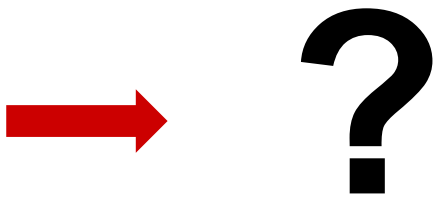
---

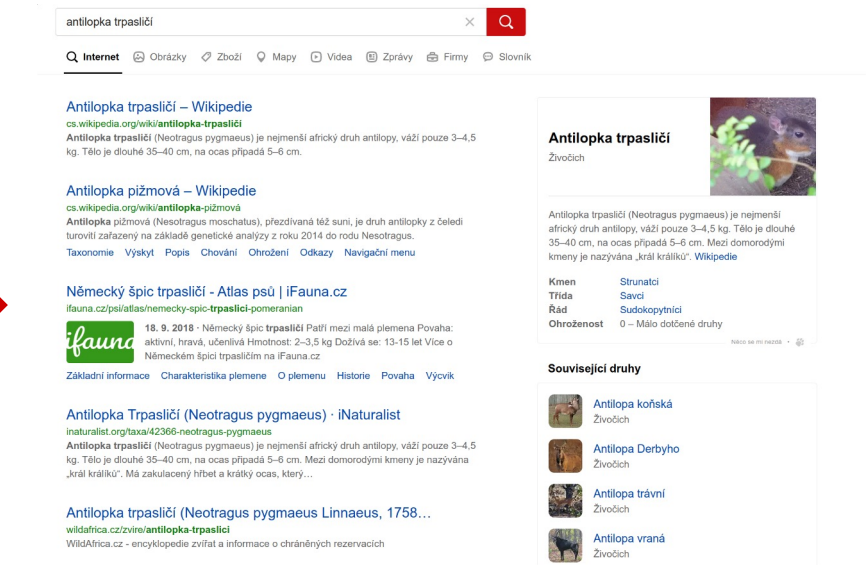
Textová média

Seznam Zprávy | Novinky.cz | Super.cz | Sport.cz | Proženy | Počasí.cz | Garáž.cz

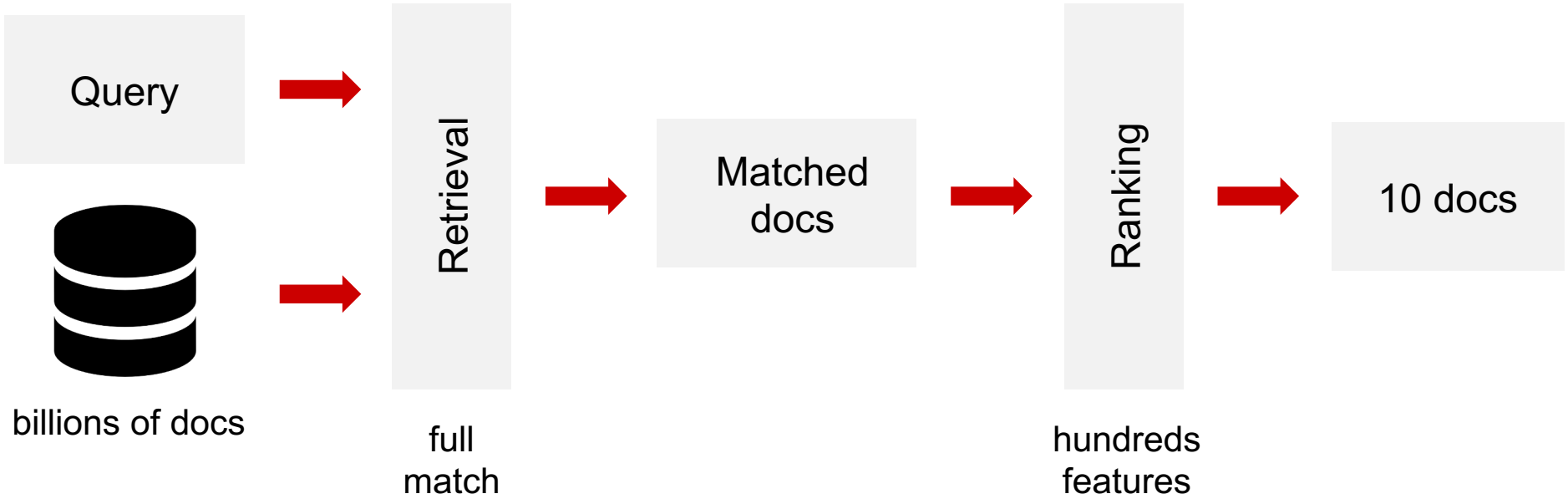
Fulltext Search of Seznam.cz



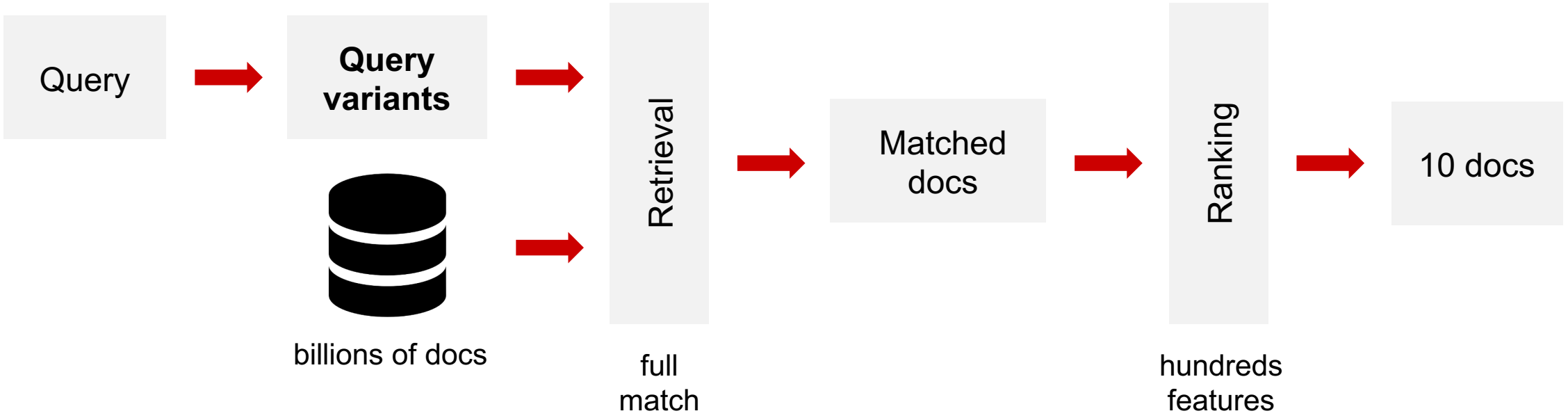




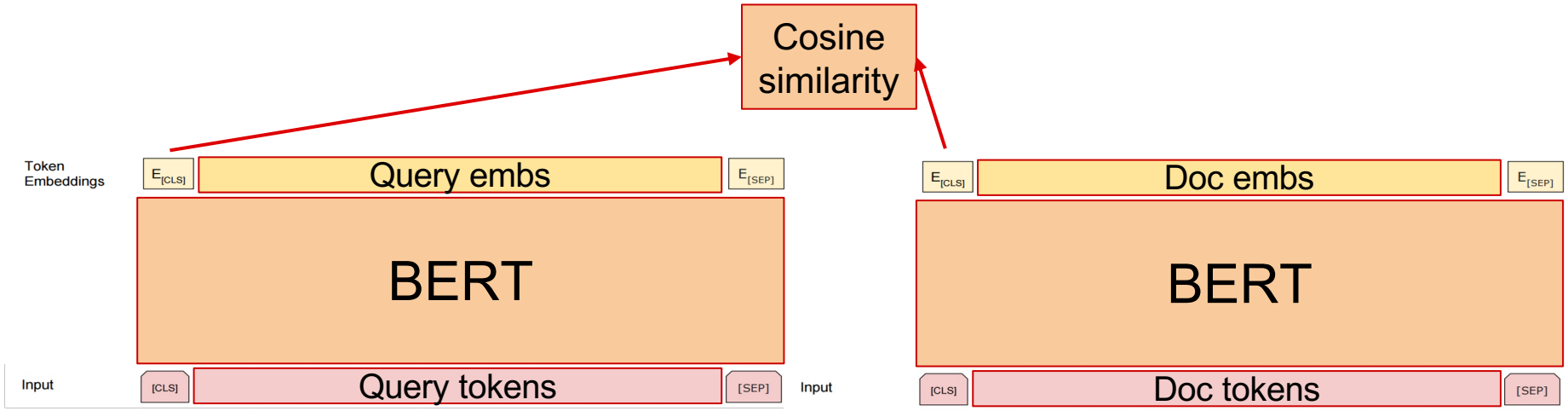
Fulltext Term Search



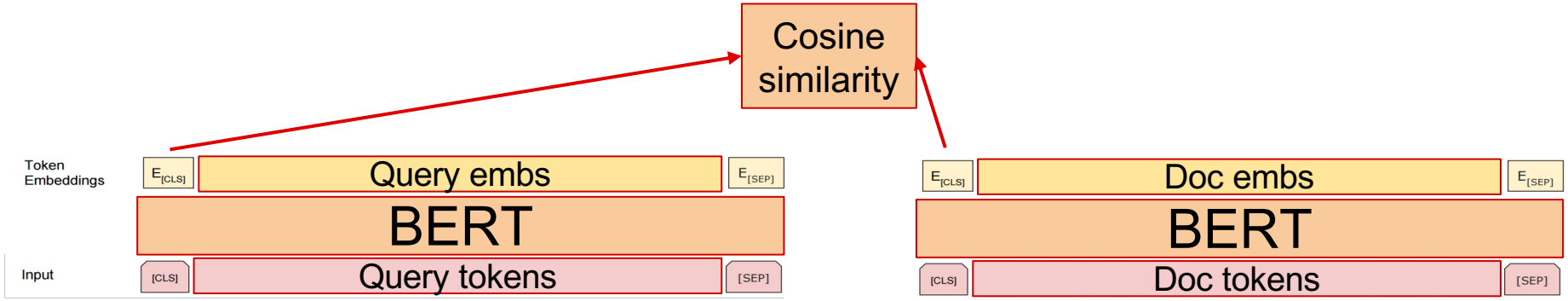
Fulltext Term Search

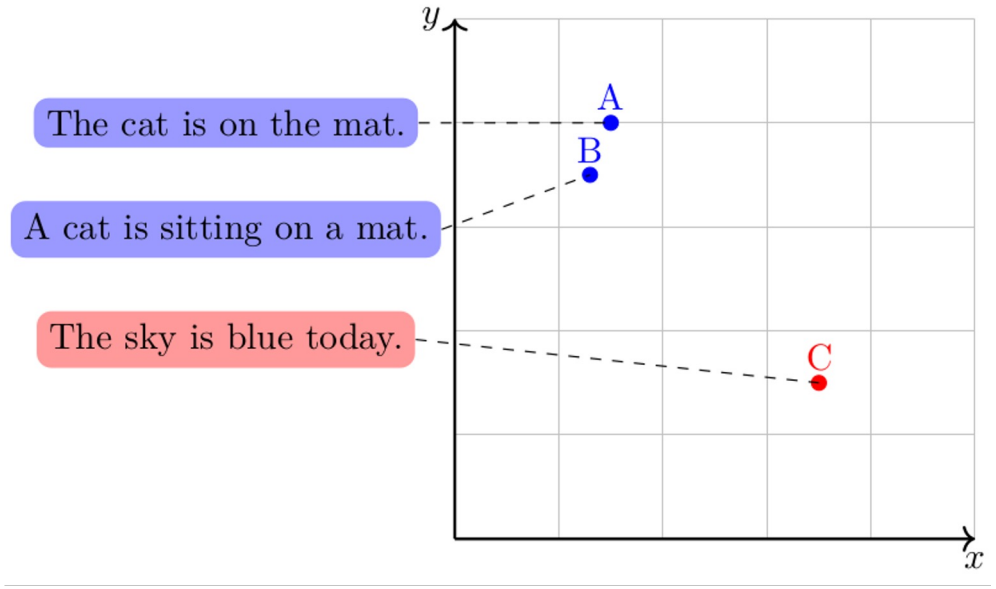


Bi-Encoder

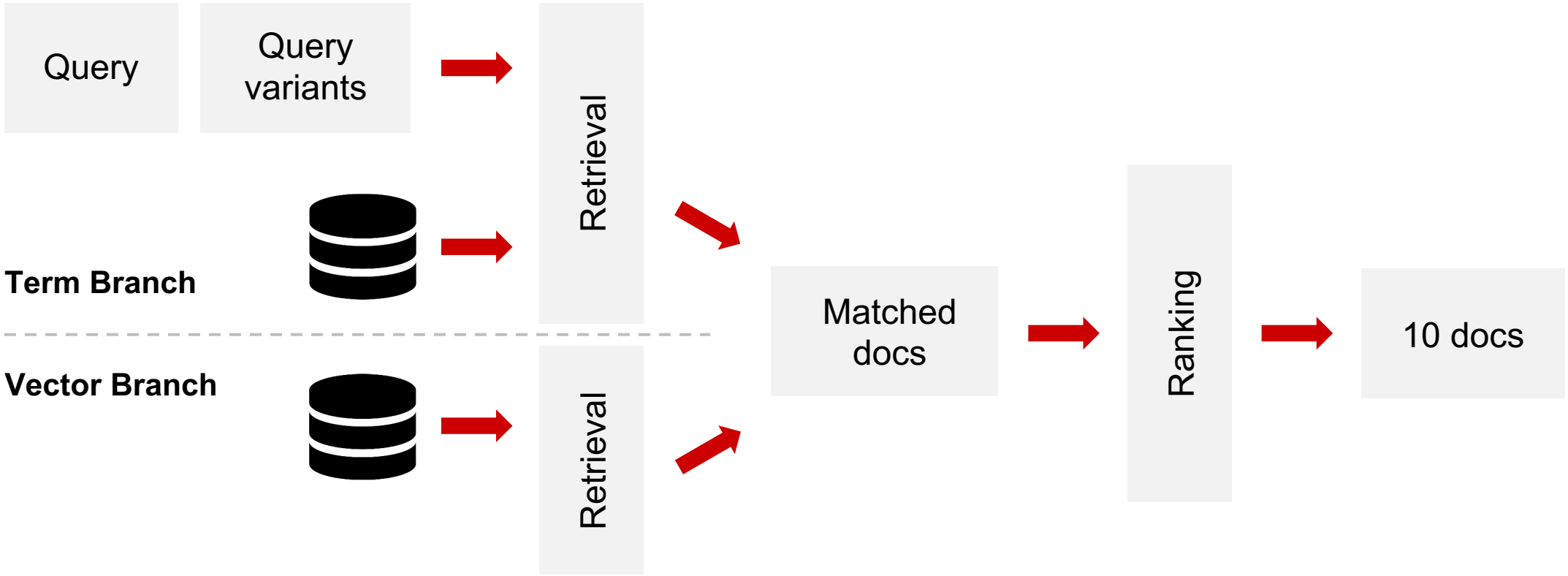

Cosine similarity
Token Embeddings
E[CLS]
Query embs
E[SEP]
BERT
Input
[CLS]
Query tokens
[SEP]
Input
E[CLS]
Doc embs
E[SEP]
BERT
[CLS]
Doc tokens
[SEP]

Bi-Encoder


Cosine similarity
Token Embeddings
E[CLS]
Query embs
E[SEP]
BERT
Input
[CLS]
Query tokens
[SEP]
E[CLS]
Doc embs
E[SEP]
BERT
[CLS]
Doc tokens
[SEP]


y
x
The cat is on the mat.
A cat is sitting on a mat.
The sky is blue today.
A
B
C

The Task


Query
Query variants
Retrieval
Term Branch
Vector Branch
Retrieval
Matched docs
Ranking
10 docs

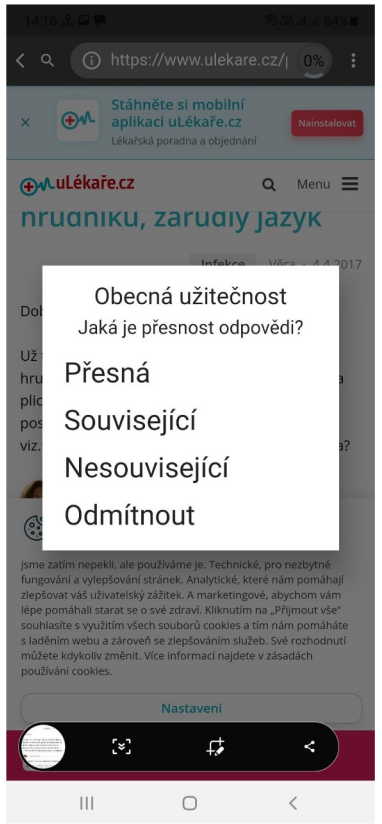
Small-E-Czech

- Electra-Small model (14M params) trained on in-house corpus of Czech web documents for 20 days
- cc-by-4.0
- more than 500k downloads

https://huggingface.co/Seznam/small-e-czech

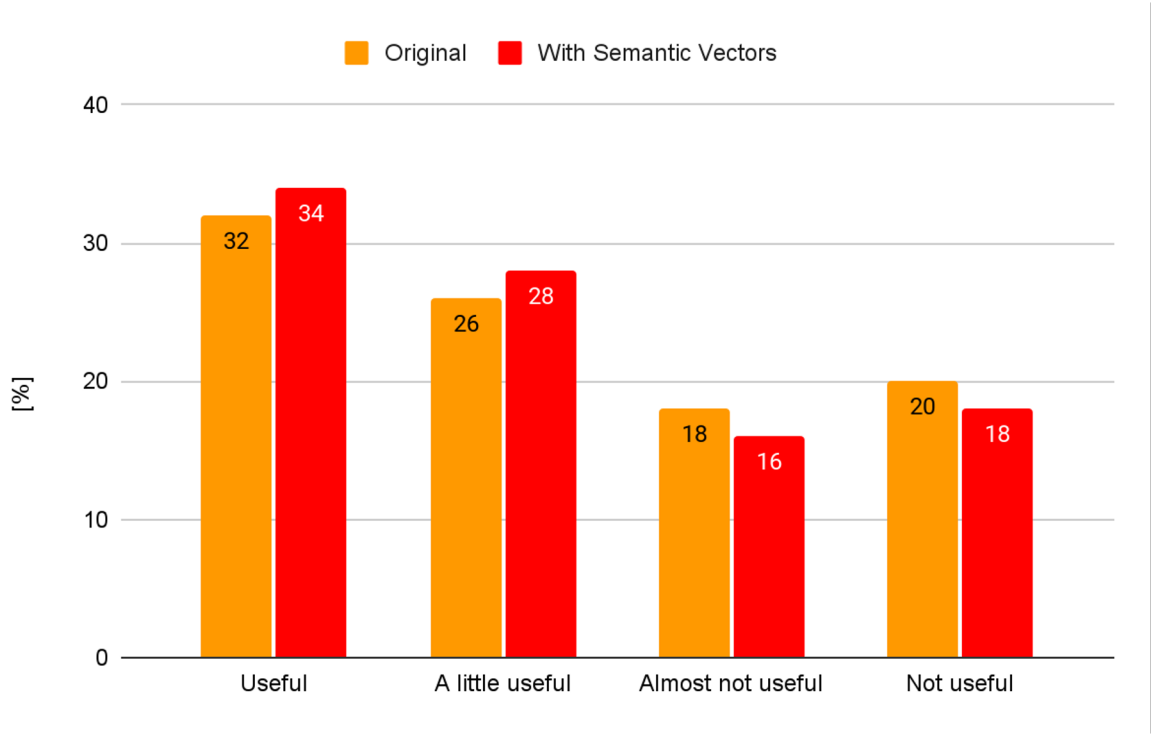
DaReCzech - Czech relevance dataset

- existing relevance annotations with 4 relevance labels:
  - Useful: (1)
  - A little useful: (0.75)
  - Almost not useful: (0.25)
  - Not useful (0)



Overall improvements

- queries sampled from traffic
- each document annotated on 4 relevance labels



SEZNAM.CZ

jak se nazýval dokument z roku 1212, který mj. zaručoval českým panovníkům

Internet | Obrázky | Zboží | Mapy | Videa | Zprávy | Firmy | Slovník

Zlatá bula sicilská – Wikipedie

cs.wikipedia.org/wiki/zlatá-bula-sicilská

Název listiny, resp. listin, je odvozen od pečeti, která je k **dokumentu** přivěšena. Fridrich II. jako král Sicílie disponoval tehdy pouze pečetí tohoto království.

Obsah prvního privilegia | Poznámky | Literatura | Související články | Externí odkazy

Přemyslovci – Wikipedie

cs.wikipedia.org/wiki/přemyslovci

Ve vedlejší (levobočné) opavské linii Přemyslovci vymřeli (po meči) až **roku** 1521.

Původ dynastie | Název | Období vlády | Poslední Přemyslovci | Tělesná charakteristika

700. výročí vymření Přemyslovců po meči Václavem III

zlate-mince.cz/crs_2006_premyslovci_info.htm

700. výročí vymření Přemyslovců po meči Václavem III, stříbrná pamětní mince **České** národní banky v hodnotě 200 Kč. Získáte ji ve specializované prodejně numismatiky v Obecním domě v Praze.

Příběhy výzkumu a vývoje: Jak rychle najít Zlatou bulu…

tripartita.cz/pribehy-vyzkumu-a-vyvoje-jak-rychle-najit-zlatou…

Tripartita, oficiálním názvem Rada hospodářské a sociální dohody **České** republiky (RHSD ČR) je společným dobrovolným dohadovacím a iniciativním orgánem odborů, zaměstnavatelů a vlády

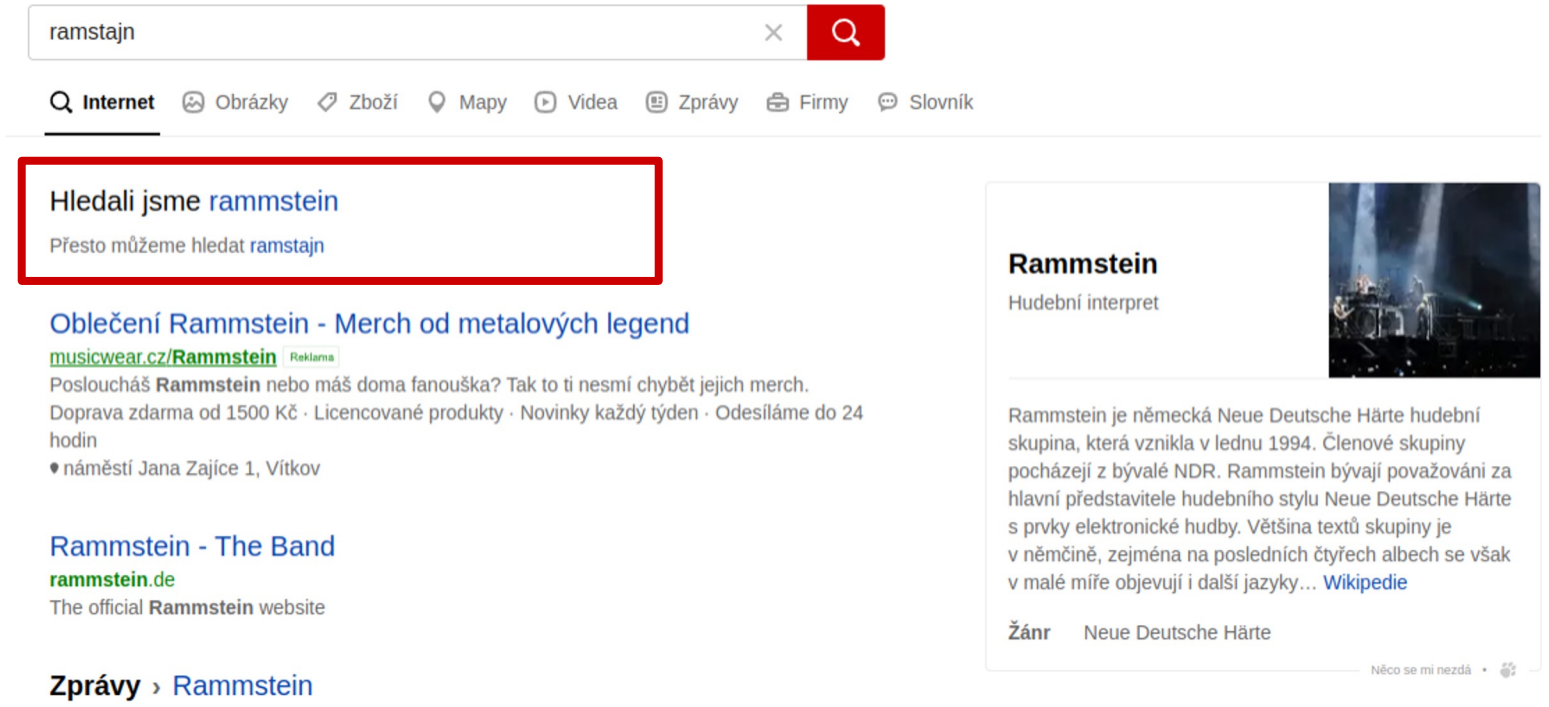
The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)

Siamese BERT-Based Model for Web Search Relevance Ranking Evaluated on a New Czech Dataset

Matěj Kocián*, Jakub Náplava*, Daniel Štancl*, Vladimír Kadlec

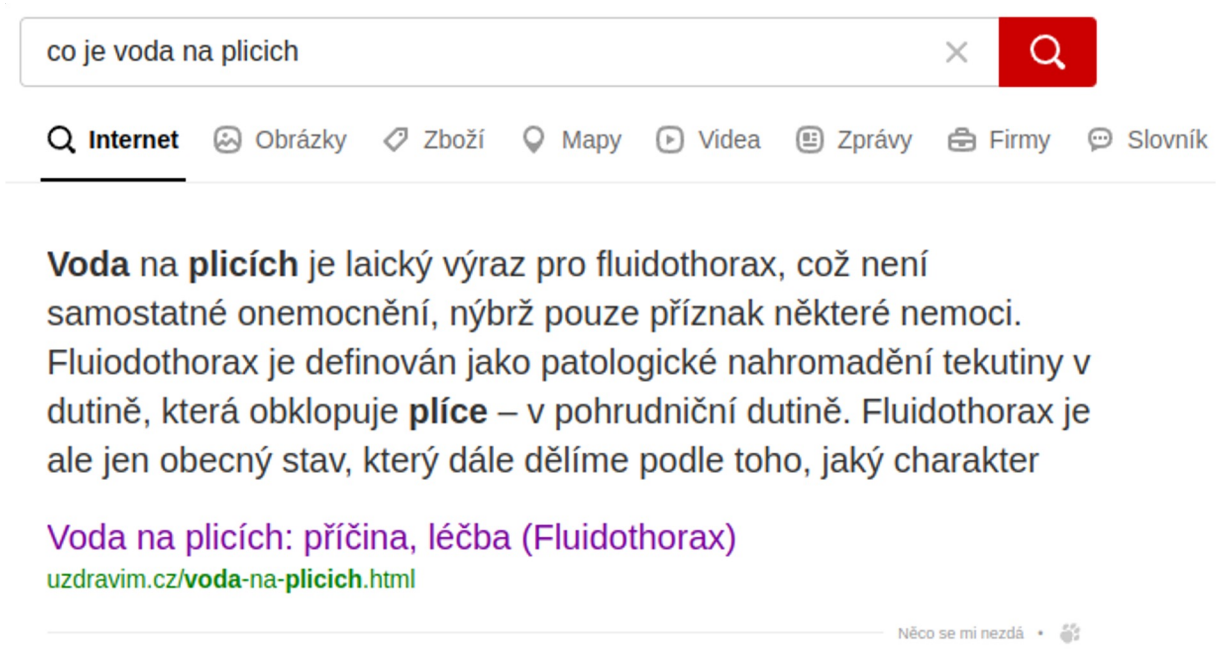
Seznam.cz, Prague, Czechia
{matej.kocian,jakub.naplava,daniel.stancl,vladimir.kadlec}@firma.seznam.cz

Query Correction



https://www.root.cz/clanky/rychla-oprava-dotazu-ve-vyhledavaci-pomoci-neuronovych-siti/

Featured Snippets



https://blog.seznam.cz/2022/01/vyhledavani-seznamu-uvadi-vlastni-uryvky-ze-stranek-neboli-featured-snippets/

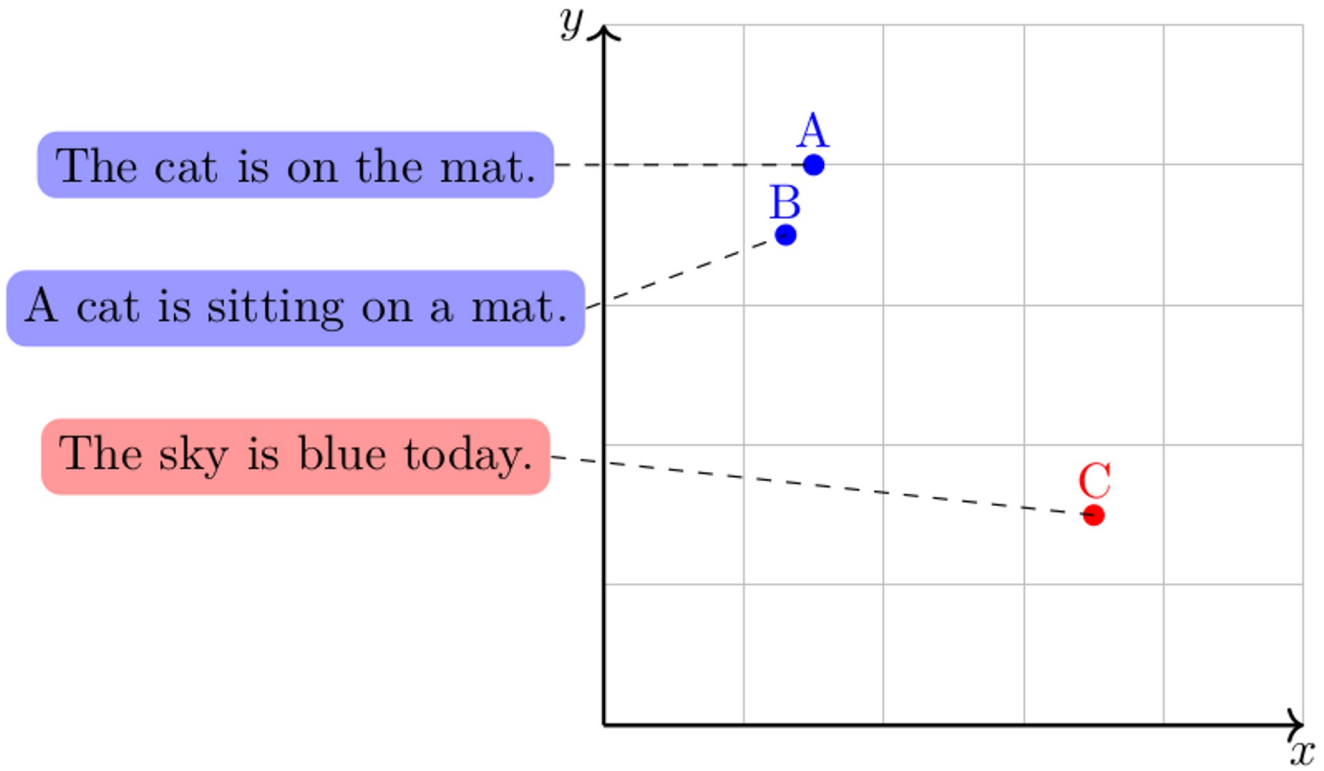
And many others

- clickbait detection (recommendation)
- review classification (firmy.cz)
- relevance (zbozi.cz)
- partial exact match (search)

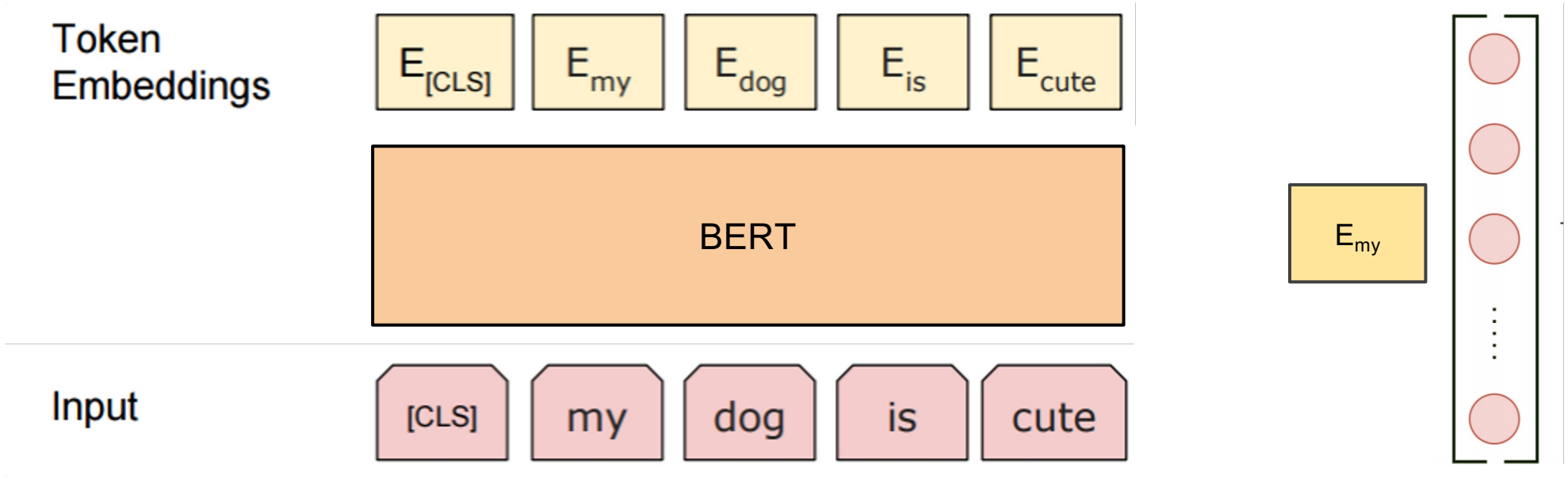
…

Czech Semantic Embedding Models

Motivation

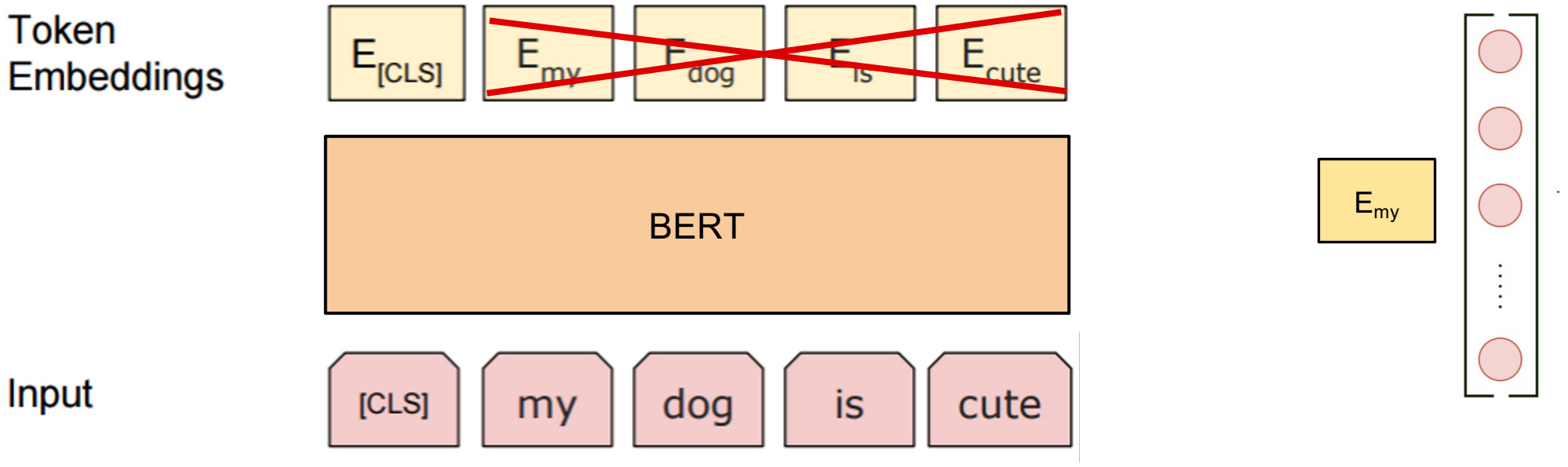

y
x
A
B
C
The cat is on the mat.
A cat is sitting on a mat.
The sky is blue today.

Motivation


Token Embeddings
E[CLS]
Emy
Edog
Eis
Ecute
BERT
Input
[CLS]
my
dog
is
cute
Emy

Motivation

- **Concentrate on sentence embeddings only (Sentence Transformers)**

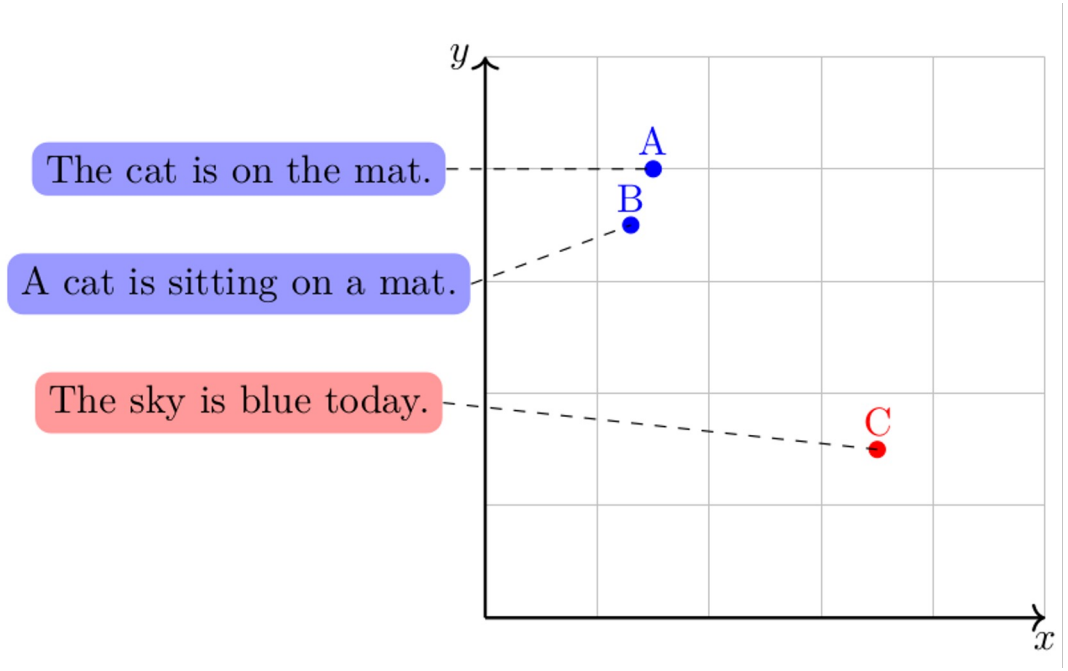


We want a small and fast Czech encoder with high-quality semantic text embedding space well suited for downstream tasks.

How to Evaluate Such Model?

2 intrinsic tasks:

- Semantic Textual Similarity
- Costra



How to Evaluate Such Model?

2 intrinsic tasks:

- Semantic Textual Similarity
- Costra

3 downstream tasks:

- Relevance (DaReCzech)
- Sentiment classification (CFD)
- Multi-label classification (CTDC)

How to Evaluate Such Model?

2 intrinsic tasks:

- Semantic Textual Similarity
- Costra

The higher the better.

3 downstream tasks:

- Relevance (DaReCzech)
- Sentiment classification (CFD)
- Multi-label classification (CTDC)

How to Train Such Model?

- Pre-training from scratch with semantic objective (RetroMAE)

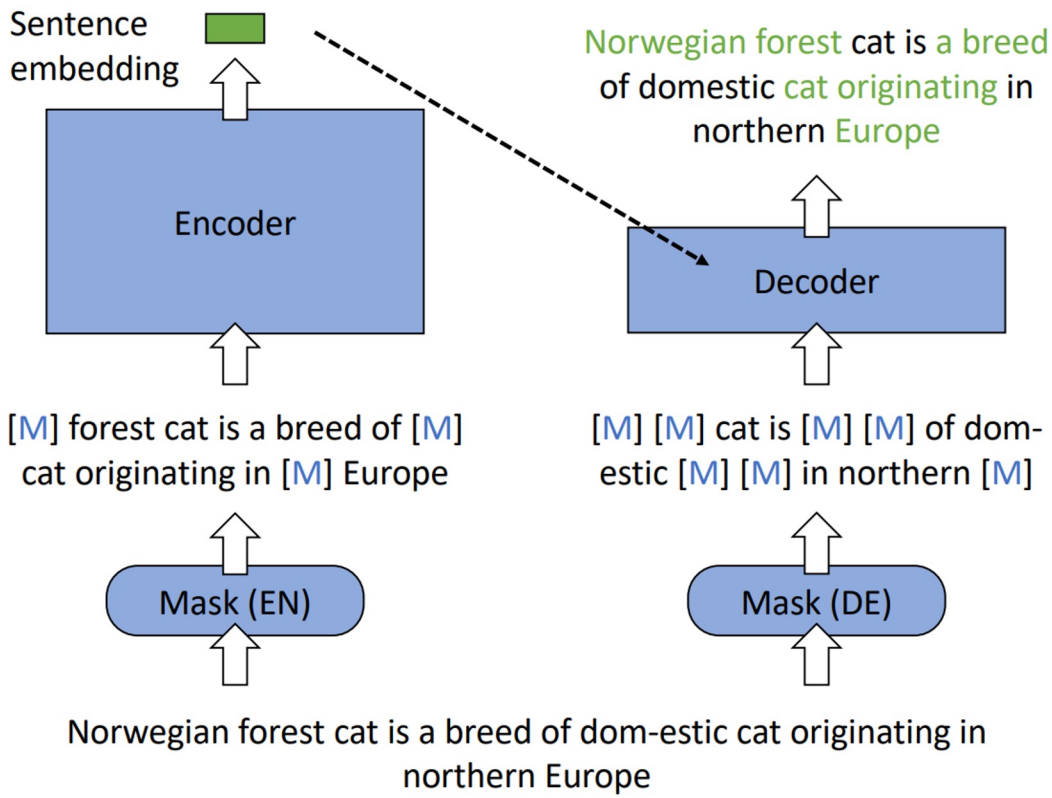


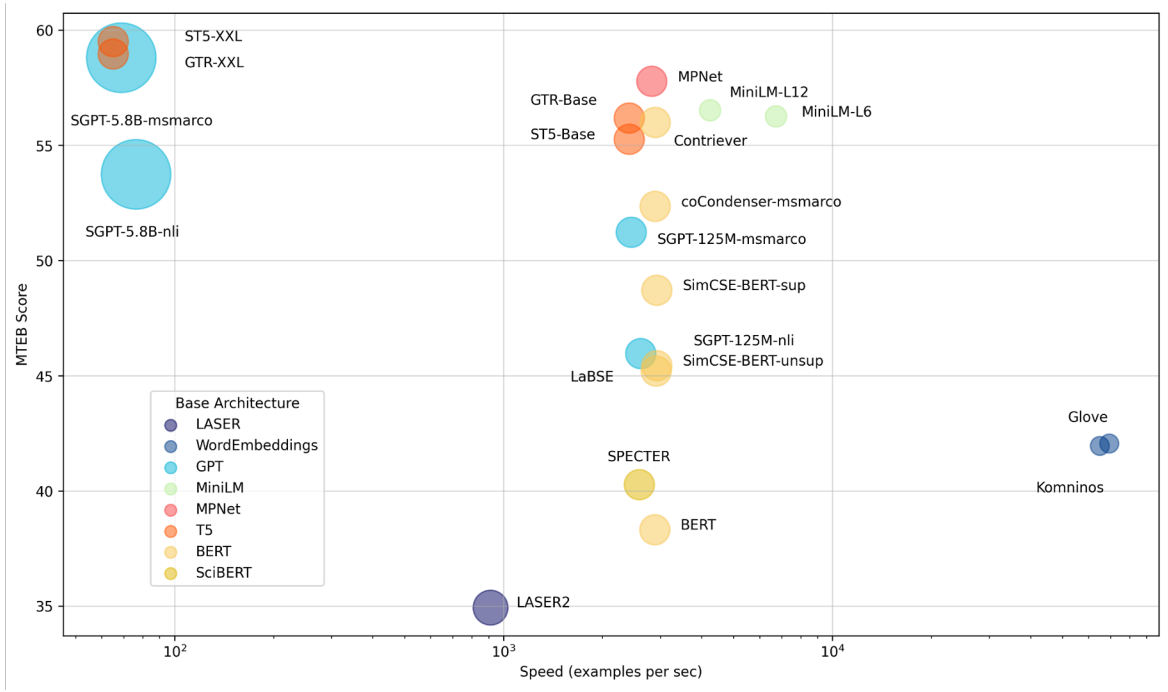
image taken from https://arxiv.org/pdf/2205.12035.pdf

How to Train Such Model?

- Pre-training from scratch with semantic objective (RetroMAE)
- Contrastive fine-tuning (not part of this presentation)
  - SIMCSE
  - RankCSE
  - INFOCSE

How to Train Such Model?

- Pre-training from scratch with semantic objective (RetroMAE)
- Contrastive fine-tuning (not part of this presentation)
- Multilingual distillation



How to Train Such Model?

- Pre-training from scratch with semantic objective (RetroMAE)
- Contrastive fine-tuning (not part of this presentation)
- Multilingual distillation

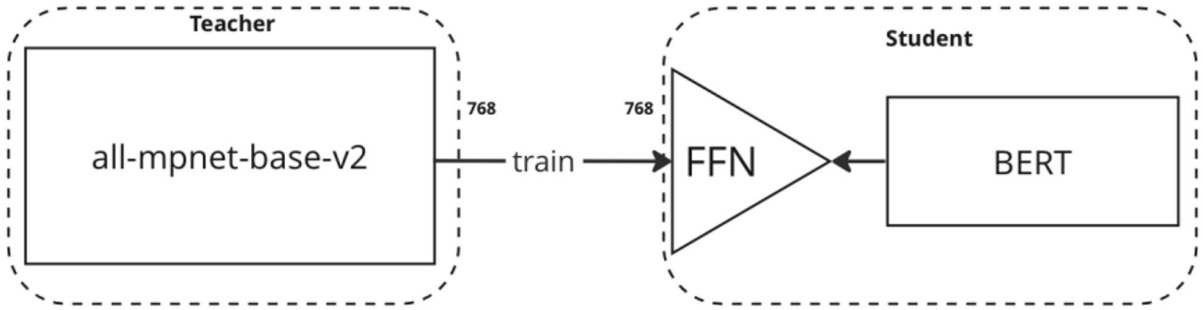
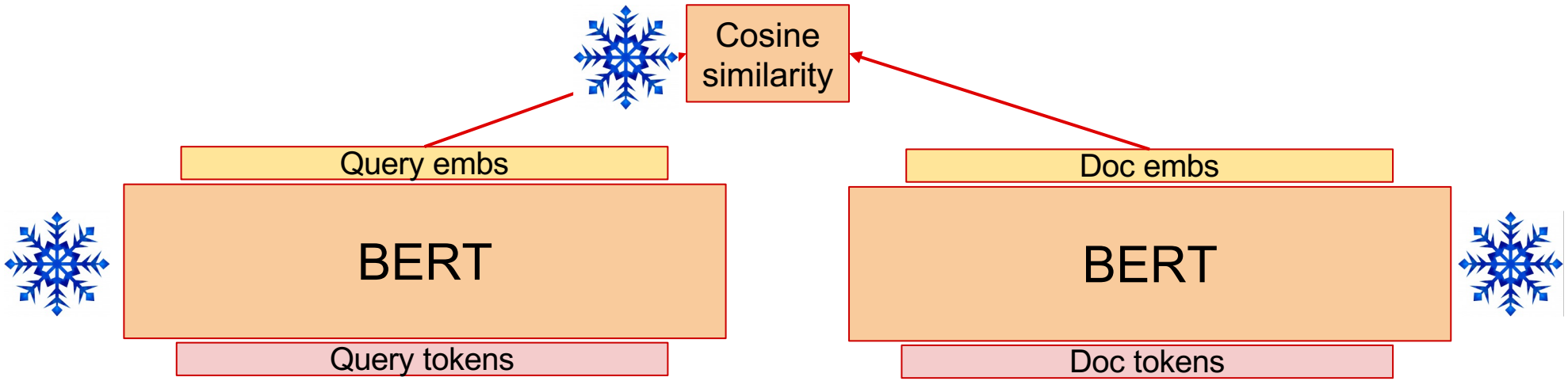


Figure 4.2 The difference in multilingual distillation using PCA and one layer of Feed-Forward Neural network (FFN). The PCA trains students on embedding size of 256, meanwhile the FFN trains the student on embedding size of 768.

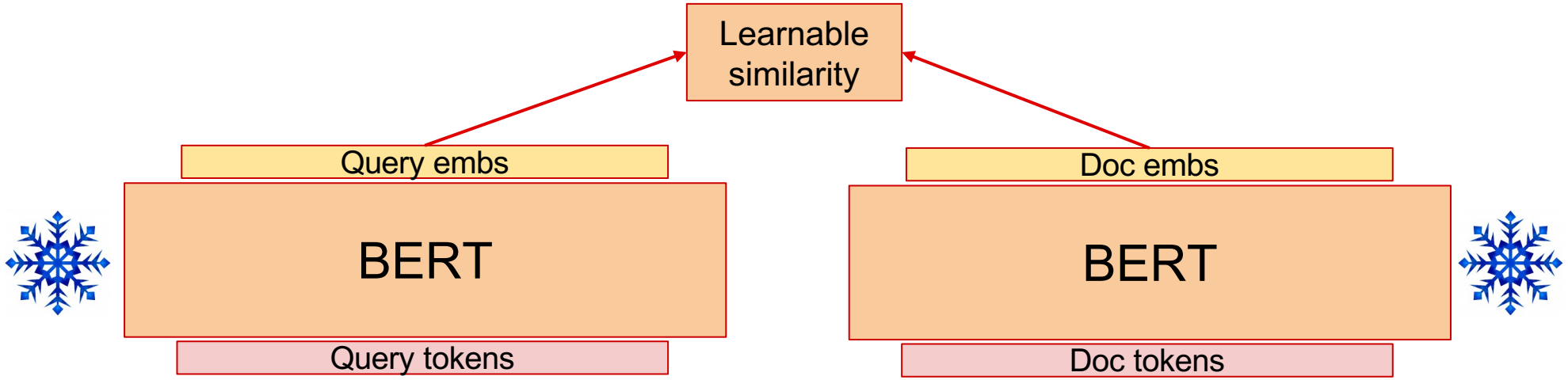
Evaluations

- Zero-shot



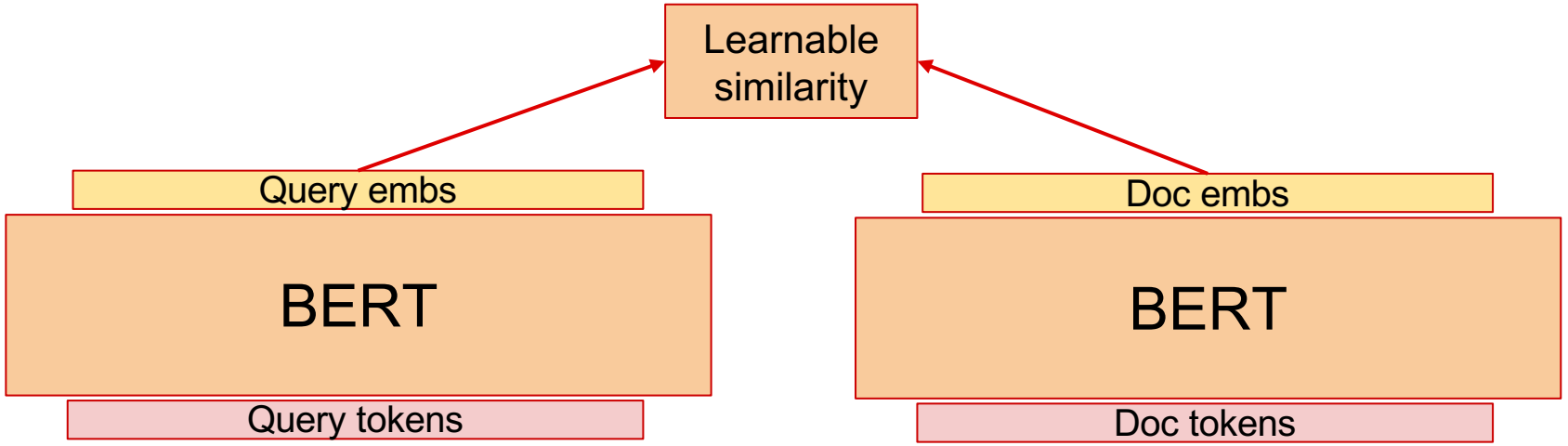
Evaluations

- Zero-shot
- Linear probing (not part of this presentation)



Evaluations

- Zero-shot
- Llinear probing (not part of this presentation)
- Full finetuning



Zero Shot

SMALL MODEL	Costra	STS	DaReCzech
Small-E-Czech	63.05	48.3	36.90 ±0.36
RetroMAE	**69.91**	**76.30**	**42.16 ±0.36**

Semantic pre-training forms better embedding space than traditional language modelling.

Zero Shot

SMALL MODEL	Costra	STS	DaReCzech
Small-E-Czech	63.05	48.3	36.90 ±0.36
RetroMAE	69.91	76.30	42.16 ±0.36
Dist-MPNet-CzEng	**71.22**	**87.60**	42.01 ±0.37
Dist-MPNet-ParaCrawl	70.42	84.25	**42.33 ±0.32**

Distillation from existing English models forms even better embedding space.

Zero Shot

MODEL	Costra	STS	DaReCzech
Small-E-Czech	63.05	48.3	36.90 ±0.36
RetroMAE	69.91	76.30	42.16 ±0.36
Dist-MPNet-CzEng	**71.22**	**87.60**	42.01 ±0.37
Dist-MPNet-ParaCrawl	70.42	84.25	**42.33 ±0.32**
OpenAiEmbeddings (ADA)	69.01	82.59	42.21 +-0.33

Our models outperform commercial alternatives.

Finetuning

SMALL MODEL	CFD	DaReCzech	CTDC
Small-E-Czech	76.94 +-1.18	43.64 +-0.37	58.12 ±1.52
RetroMAE	76.85 +-1.16	45.29 +-0.34	84.58 ±0.37
Dist-MPNet-CzEng	**78.73 +-1.39**	**45.75 +-0.34**	85.85 ±0.21
Dist-MPNet-ParaCrawl	77.42 +-1.60	45.55 +-0.33	**86.02 ±0.12**

Zero-shot observations hold also when fine-tuning whole model.

Comparison to BASE-sized models

SMALL MODEL	Costra	STS	CFD	DaReCzech	CTDC
LaBSE	70.63	82.91	79.79 +-1.07	46.15 +-0.34	87.97 ±0.31
Czert-b-base-cased	**72.08**	74.79	78.73 +-1.25	45.63 +-0.34	88.69 ±0.21
FERNET-C5	67.57	65.46	**82.00 +-1.43**	**45.87 +-0.34**	**89.56 ±0.19**
RobeCzech	63.94	69.41	80.54 +-0.93	45.54 +-0.31	86.01 +-0.24
Dist-MPNet-CzEng	71.22	**87.60**	78.73 +-1.39	45.75 +-0.34	85.85 ±0.21
Dist-MPNet-ParaCrawl	70.42	84.25	77.42 +-1.60	45.55 +-0.33	86.02 ±0.12

BASE | SMALL

Small-sized models are competitive to their BASE counterparts.

Query: **Jaký je nejlepší způsob, jak odstranit skvrny z oblečení?**

Most Similar Query	Rank	Similarity Score
jak odstranit skvrny od oblečení	1	0.964
jak vyčistit skvrny na oblečení	5	0.918
jak odstranit mastný flek z oblečení	10	0.902
odstranit mastné skvrny z prádla	50	0.823
jak odstranit skvrny od fixy	100	0.798

Internal Applications

Organic Search:

- Results for complex queries improved
- 2 % improvement [P@10]

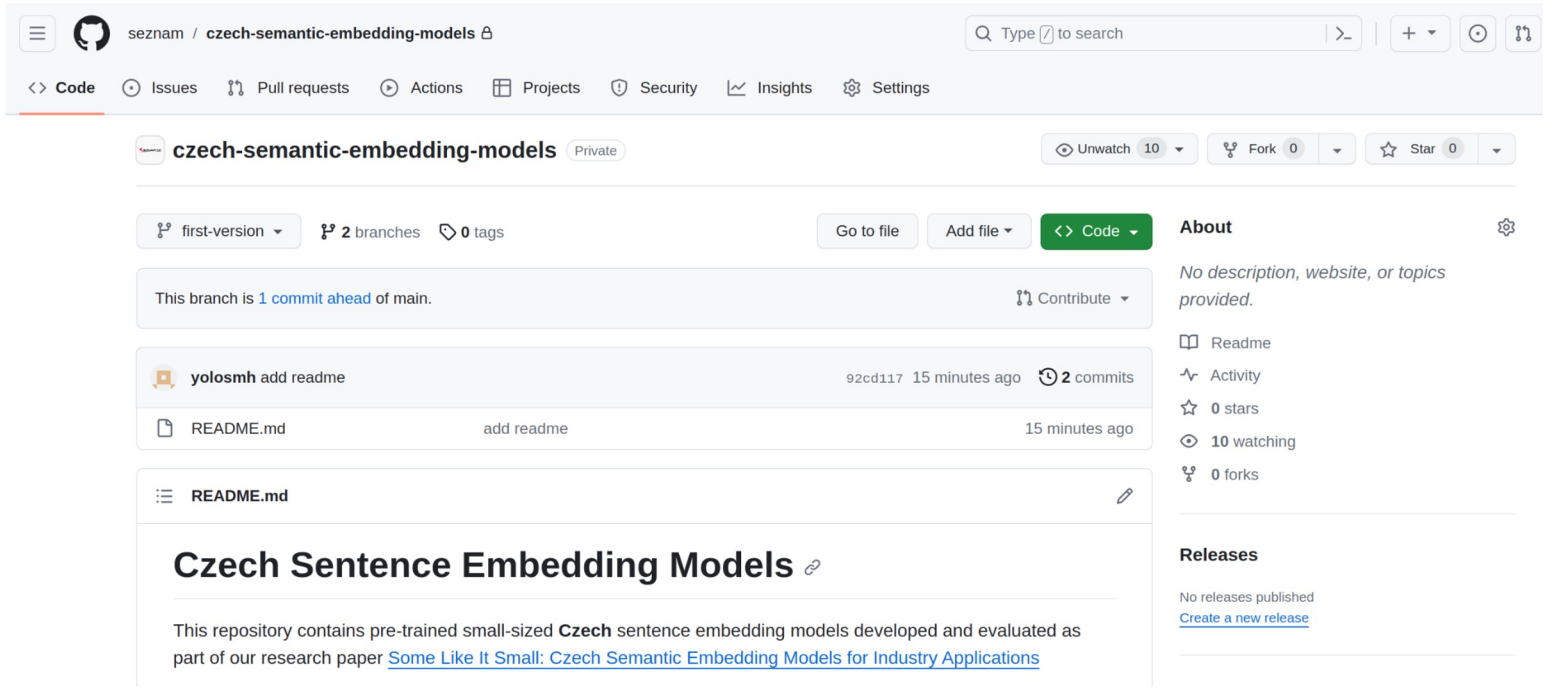
Featured Snippets:

- Error rate reduced by 20 %

Image Search:

- Relative NDCG error rate reduced by 3.2%
- 7 % more relevant images retrieved

Models Available For You



https://github.com/seznam/czech-semantic-embedding-models

Some Like It Small: Czech Semantic Embedding Models for Industry Applications

Jiří Bednář, Jakub Náplava, Petra Barančíková, Ondřej Lisický

Seznam.cz, Prague, Czech Republic
{jiri.bednar,jakub.naplava,petra.barancikova,ondrej.lisicky}@firma.seznam.cz

Future Plans

Future Plans

- bigger models

Future Plans

- bigger models
- combination with generative LLMs - pre-selection of relevant context

Future Plans

- bigger models
- combination with generative LLMs - pre-selection of relevant context

Závěrem

Jakub Náplava

ML Research Team Lead

jakub.naplava@firma.seznam.cz

Prostor pro dotazy