



Encoder-based Jazykové Modely v Seznamu

Jakub Náplava

Náš život s velkými i malými jazykovými modely



Seznam.cz

SEZNAM.CZ

S | Televize Seznam

stream

SREALITY.CZ

SPORT.CZ


EXPRESFM

EMAIL

S DOVOLENÁ.CZ

S BLOG

SKLIK.CZ

S AUTO.CZ

S MOTO.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/

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

erová

bit draka, ale nevíte jak? Pomůžeme vám
y nejlepší návody, jak si vyrobit draka a užít
ou rodinou.

Přihlásit

Založit nový účet

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

Sřelba při zásahu URNA v Praze, zraněný mladík skončil v umělém spánku

Zemřel herec Otmar Brancuzský

Reklama · Koupit

Výprodej povlečení, RŮZNÉ ROZMĚRY A VZORY

14:19

SEZNAM.CZ

Kde je luhansk

Q Internet Obrázky Zboží Mapy

Luhansk (Луганськ)
Město

gorod-lugansk.com

Luhanská oblast, Ukrajina

Luhansk či Lugansk je město ležící na východní Ukrajině na soutoku řek Vichivka a Luhanka, které se nedaleko vlévají do Severního Donce. Luhansk je administrativním centrem průmyslově... Číst dále

Počasi Podrobná předpověď

Dnes Úterý Středa Čtvrtek Pátek

Ohodnoťte výsledek hledání

Contents

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



Encoder based Language Models in Seznam.cz

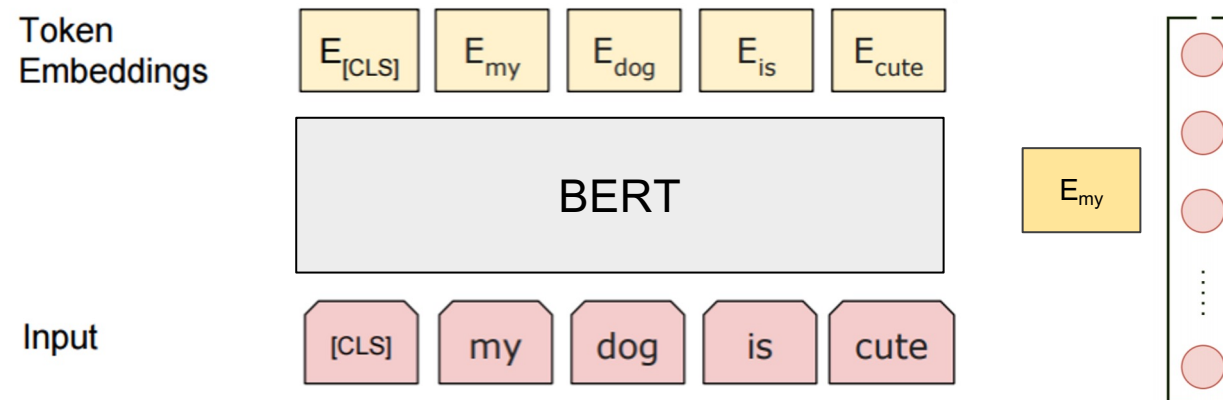


What are Encoder Language Models?

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

 SDOVOLEJÁ.CZ

volná místa

Textová média

Seznam Zprávy |

Novinky.cz

 SUPER.CZ

 SPORT.CZ

 PROŽENY

 POČASÍ.CZ

 GARÁŽ.CZ



Fulltext Search of Seznam.cz

antilopka trpasličí

- antilopka trpasličí
- antikvariát
- antikvariát praha
- antigenní testy
- antigenní test
- antik maiselova



antilopka trpasličí

Q Internet Oobrázky Zboží Mapy Vídea Zprávy Firmy Slovník

Antilopka trpasličí – Wikipedie
cs.wikipedia.org/wiki/antilopka-trpaslicí
Antilopka trpasličí (Neotragus pygmaeus) je nejmenší africký druh antilopy, váží pouze 3–4,5 kg. Tělo je dlouhé 35–40 cm, na ocas připadá 5–6 cm.

Antilopka pížmová – Wikipedie
cs.wikipedia.org/wiki/antilopka-pizmová
Antilopka pížmová (Neotragus moschatus), přezdívaná též suni, je druh antilopy z čeledi turovití zařazený na základě genetické analýzy z roku 2014 do rodu Neotragus.
Taxonomie Výskyt Popis Chování Ohrožení Odkazy Navigační menu

Německý špic trpasličí - Atlas psů | iFauna.cz
ifauna.cz/psi/atlas/nemecky-spice-trpaslici-pomeranian
 18. 9. 2018 · Německý špic trpasličí Patří mezi malá plemena Povaha: aktivní, hravá, učenlivá Hmotnost: 2–3,5 kg Dožívá se: 13-15 let Více o Německém špic trpasličím na iFauna.cz
Základní informace Charakteristika plemene O plemenu Historie Povaha Výcvik

Antilopka Trpasličí (Neotragus pygmaeus) · iNaturalist
inaturalist.org/taxa/42366-neotragus-pygmaeus
Antilopka trpasličí (Neotragus pygmaeus) je nejmenší africký druh antilopy, váží pouze 3–4,5 kg. Tělo je dlouhé 35–40 cm, na ocas připadá 5–6 cm. Mezi domorodými kmeny je nazývána „král králiků“. Má zakulacený hrbel a krátký ocas, který...

Antilopka trpasličí (Neotragus pygmaeus Linnaeus, 1758...
wildafrica.cz/zvire/antilopka-trpaslici
WildAfrica.cz - encyklopedie zvířat a informace o chráněných rezervacích

Antilopka trpasličí
Zvivočich

Antilopka trpasličí (Neotragus pygmaeus) je nejmenší africký druh antilopy, váží pouze 3–4,5 kg. Tělo je dlouhé 35–40 cm, na ocas připadá 5–6 cm. Mezi domorodými kmeny je nazývána „král králiků“. Wikipedie

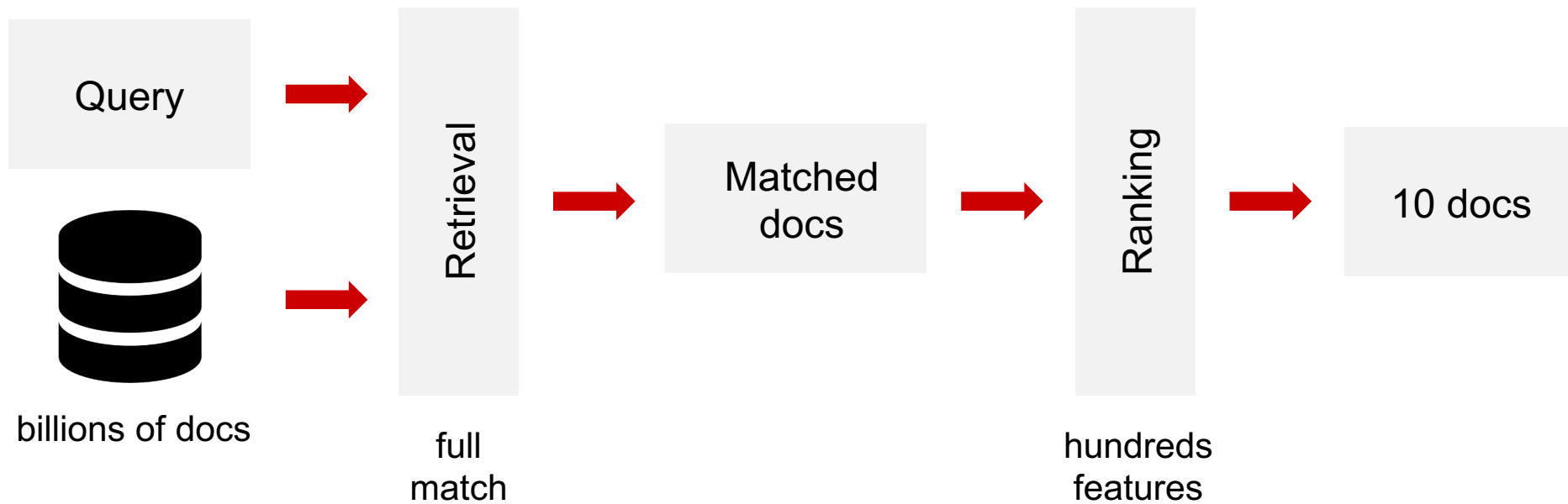
Kmen	Stunatci
Třída	Savci
Rád	Sudokopytníci
Ohroženost	0 – Málo dotčené druhy

Související druhy

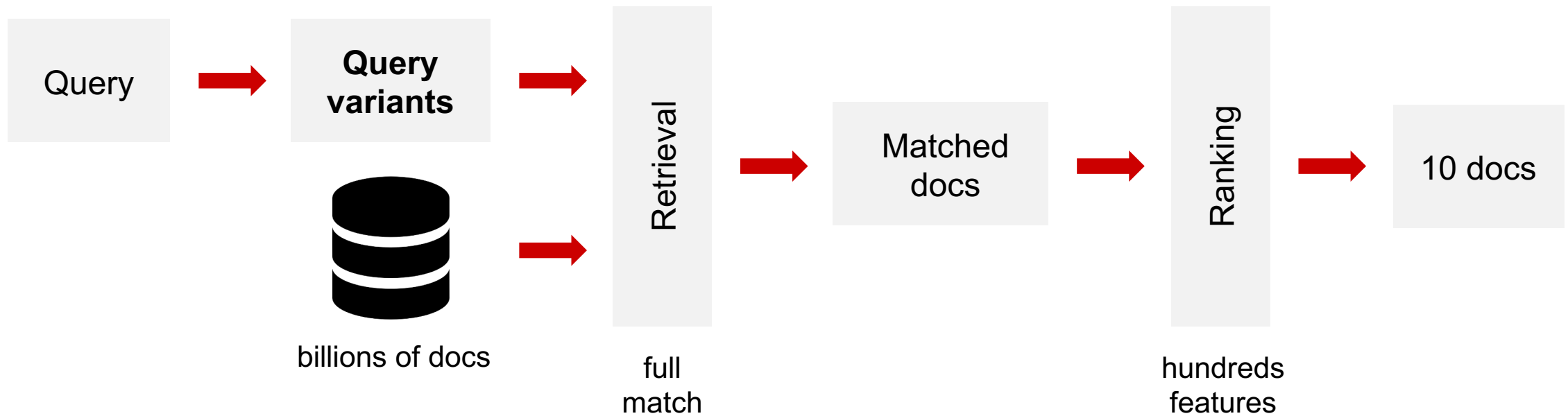
- Antilopa koňská Zvivočich
- Antilopa Derbyho Zvivočich
- Antilopa trávni Zvivočich
- Antilopa vraná Zvivočich



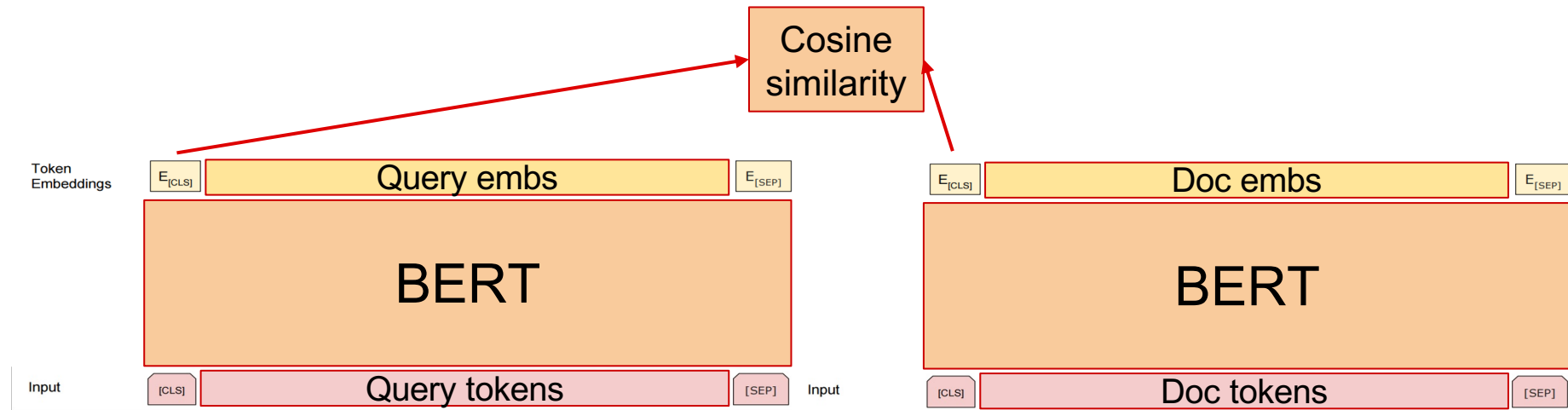
Fulltext Term Search



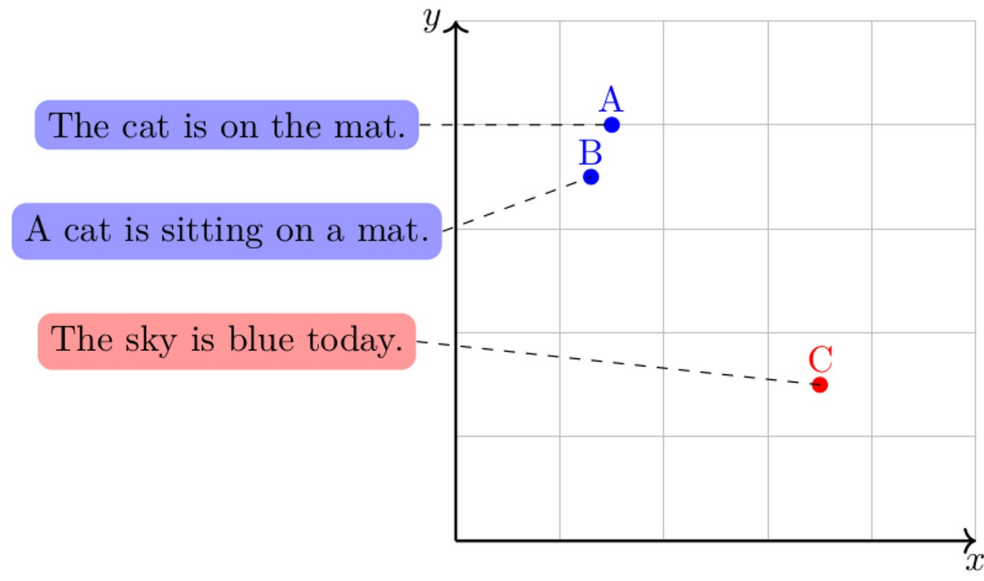
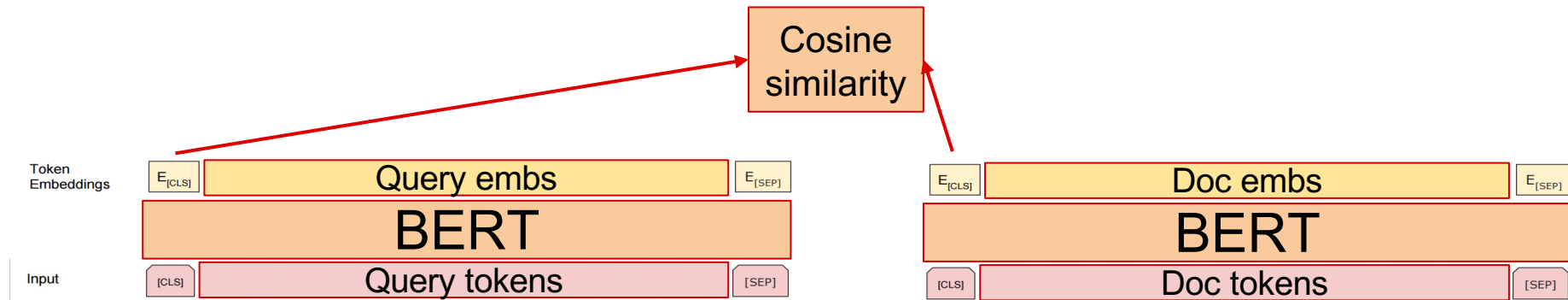
Fulltext Term Search



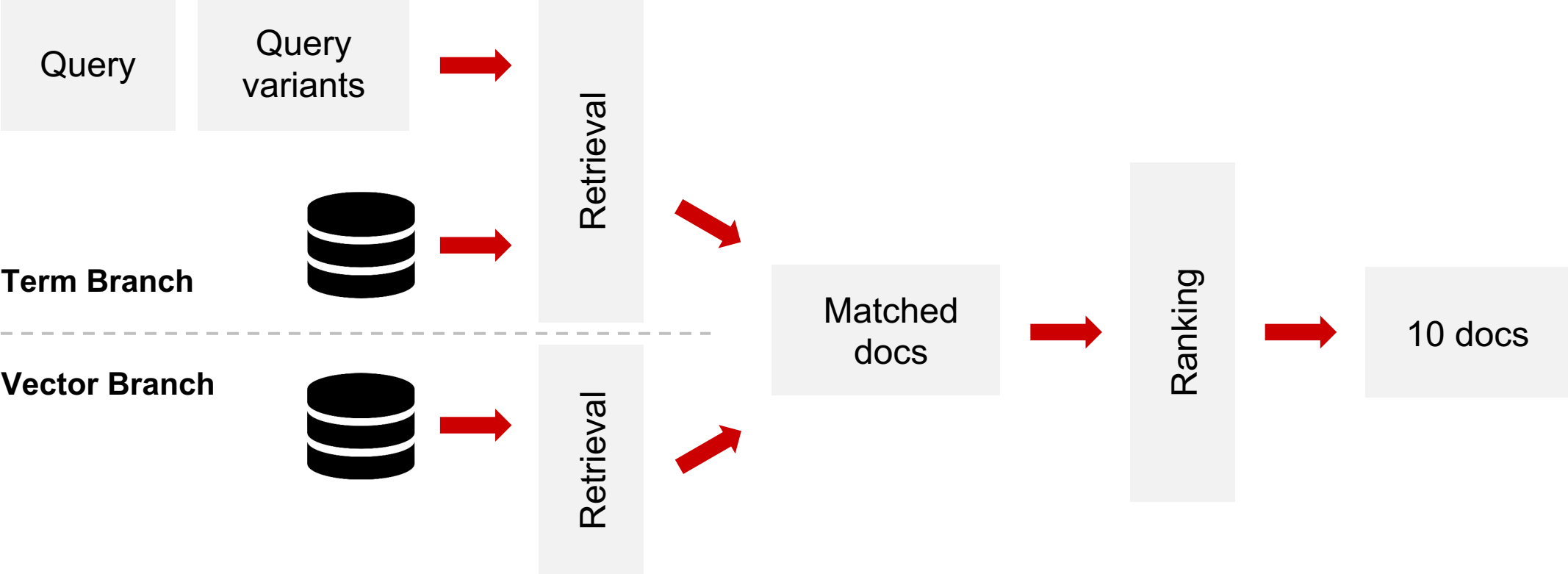
Bi-Encoder



Bi-Encoder



The Task



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



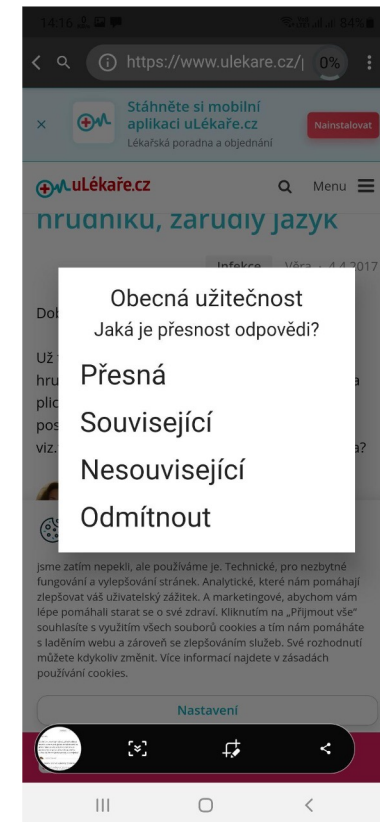
<https://huggingface.co/Seznam/small-e-czech>



DaReCzech - Czech relevance dataset

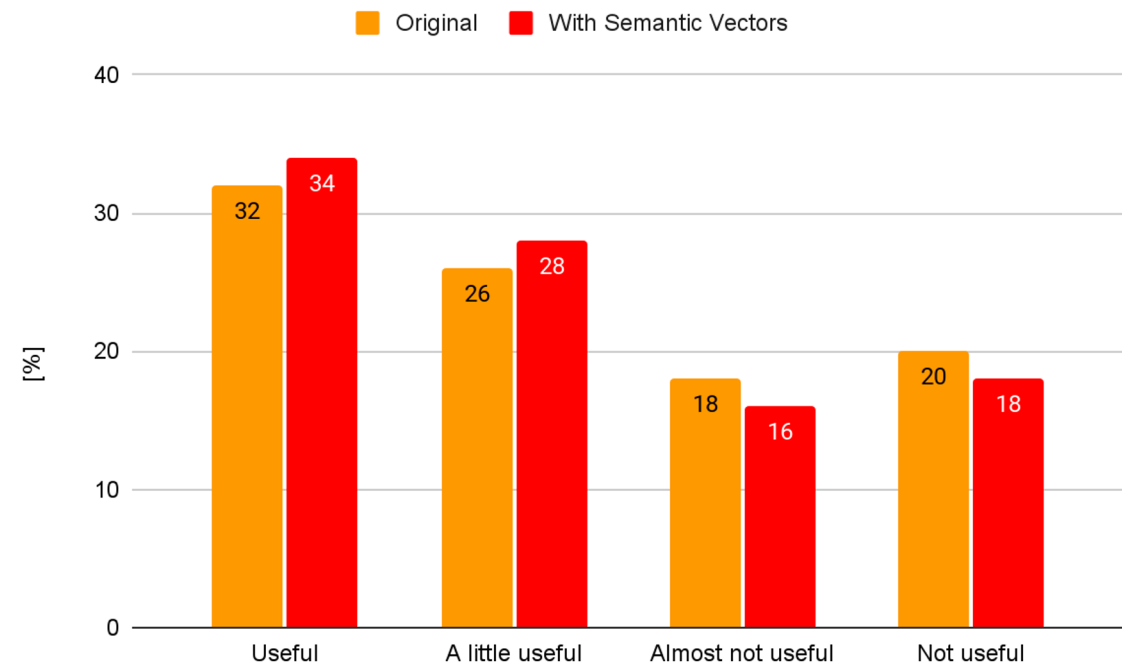
- existing relevance annotations with 4 relevance labels:

- 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





Zlatá bula sicilská – Wikipedie

[cs.wikipedia.org/wiki/zlatá-bula-sicilská](https://cs.wikipedia.org/wiki/zlat%C3%A1-bula-sicilsk%C3%A1)

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

Search bar: ramstajn

Navigation: Internet, Obrázky, Zboží, Mapy, Vídea, Zprávy, Firmy, Slovník


Hledali jsme **rammstein**
Přesto můžeme hledat **ramstajn**

Oblečení Rammstein - Merch od metalových legend
musicwear.cz/Rammstein Reklama
Posloucháš **Rammstein** nebo máš doma fanouška? Tak to ti nesmí chybět jejich merch.
Doprava zdarma od 1500 Kč · Licencované produkty · Novinky každý týden · Odesíláme do 24 hodin
• náměstí Jana Zajíce 1, Vítkov

Rammstein - The Band
rammstein.de
The official **Rammstein** website

Zprávy › [Rammstein](#)

Rammstein
Hudební interpret



Rammstein je německá Neue Deutsche Härte hudební skupina, která vznikla v lednu 1994. Členové skupiny pocházejí z bývalé NDR. Rammstein bývají považováni za hlavní představitele hudebního stylu Neue Deutsche Härte s prvky elektronické hudby. Většina textů skupiny je v němčině, zejména na posledních čtyřech albech se však v malé míře objevují i další jazyky... [Wikipedie](#)

Žánr Neue Deutsche Härte

Něco se mi nezdá



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



Featured Snippets

co je voda na plicích × 🔍

🔍 Internet 🖼️ Obrázky 🛒 Zboží 📍 Mapy 📺 Vídea 📧 Zprávy 🏢 Firmy 💬 Slovník

Voda na plicích je laický výraz pro fluidothorax, což není samostatné onemocnění, nýbrž pouze příznak některé nemoci. Fluidothorax je definován jako patologické nahromadění tekutiny v dutině, která obklopuje **plic** – v pohrudniční dutině. Fluidothorax je ale jen obecný stav, který dále dělíme podle toho, jaký charakter

Voda na plicích: příčina, léčba (Fluidothorax)

uzdravim.cz/voda-na-plicich.html

Něco se mi nezdá • 🐾



<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)

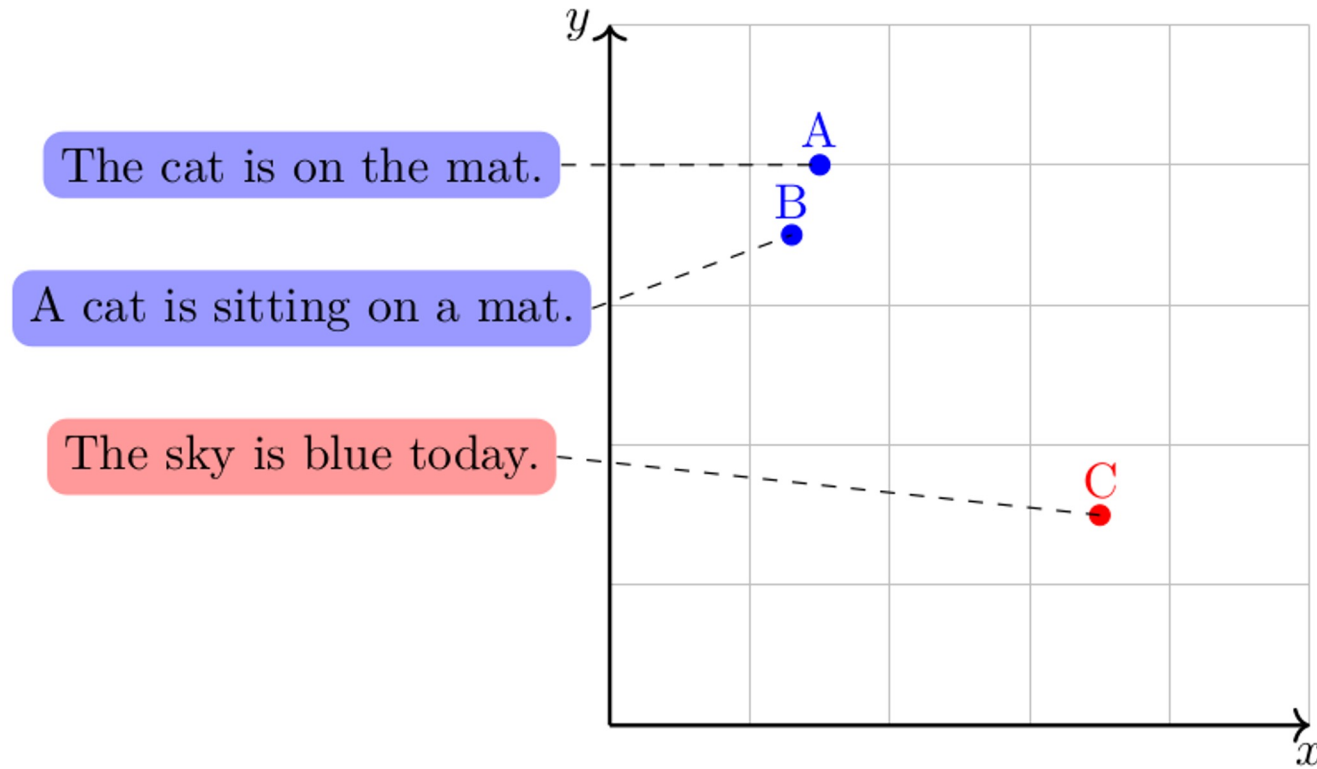
...



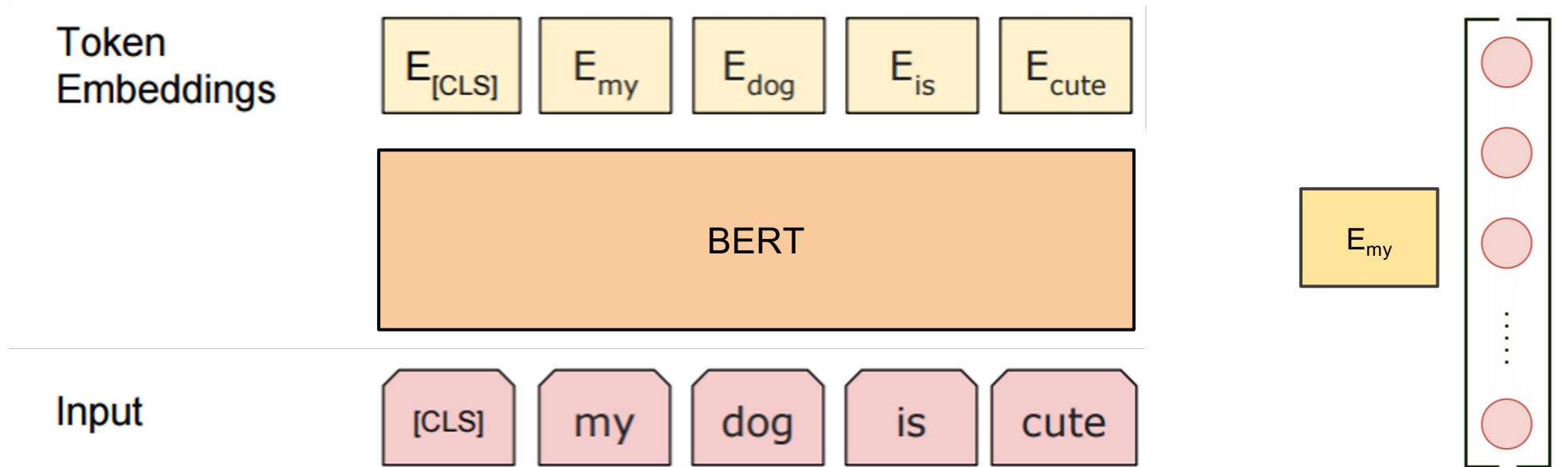
Czech Semantic Embedding Models



Motivation

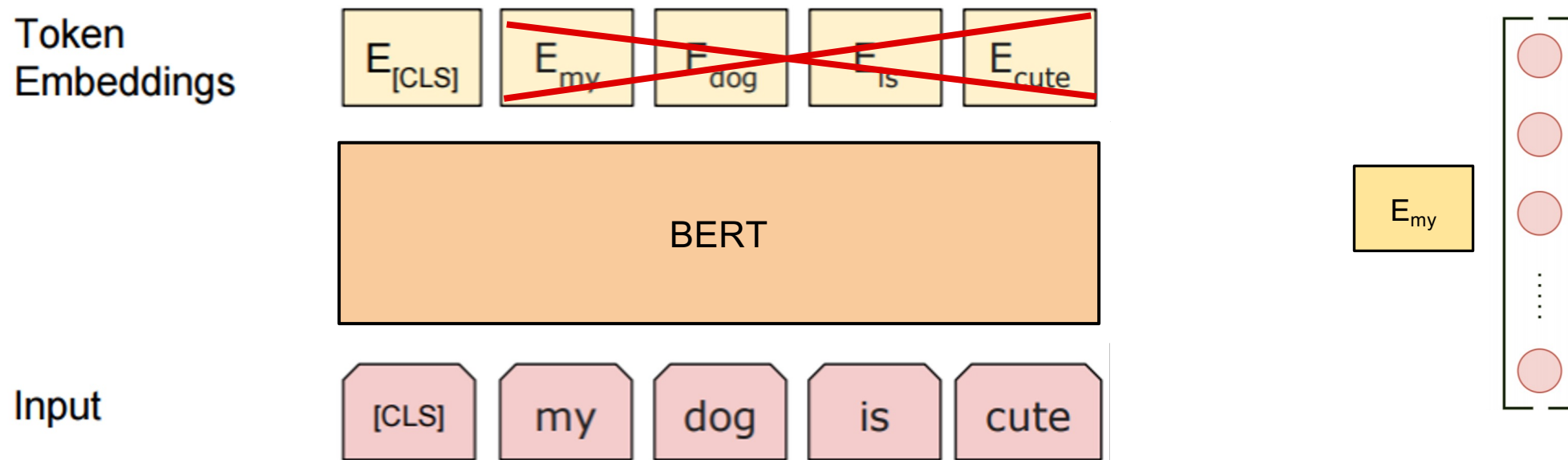


Motivation



Motivation

- ✓ Concentrate on sentence embeddings only (Sentence Transformers)

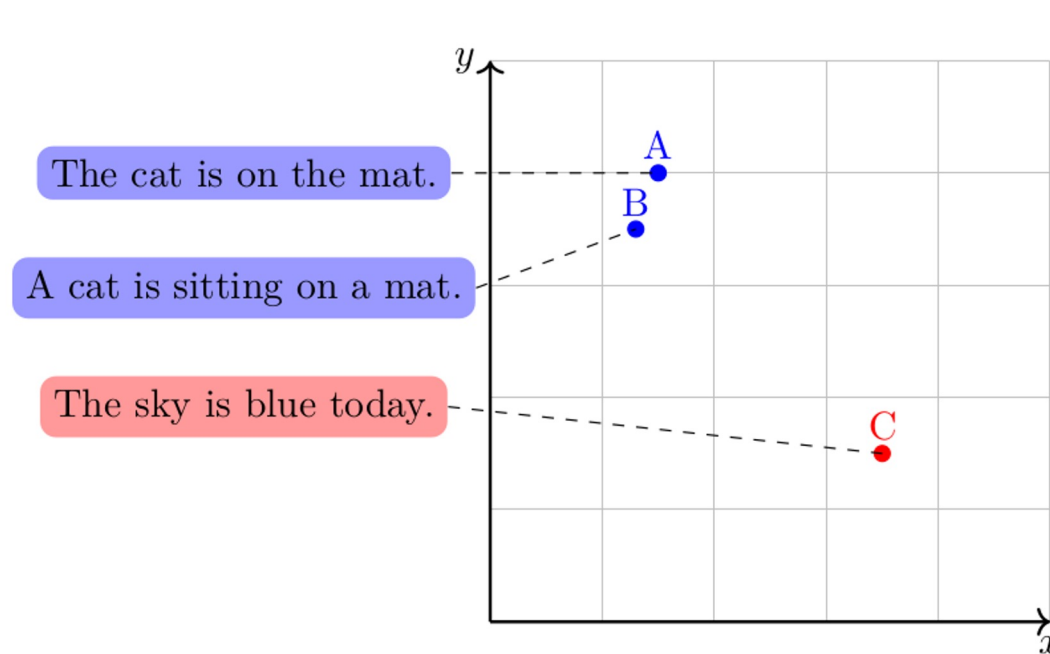


We want a small and fast Czech encoder with high-quality semantic text embedding space well suited for downstream tasks.

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

3 downstream tasks:

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

The higher the better.



How to Train Such Model?

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

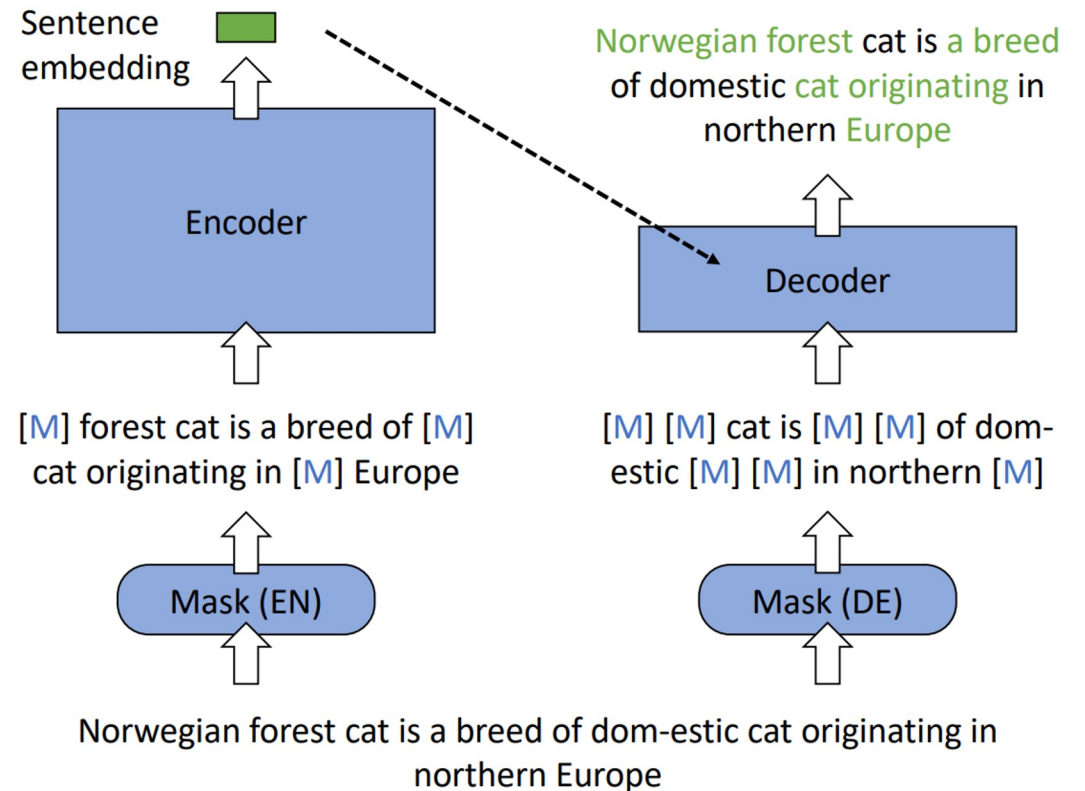


image taken from <https://arxiv.org/pdf/2205.12035.pdf>



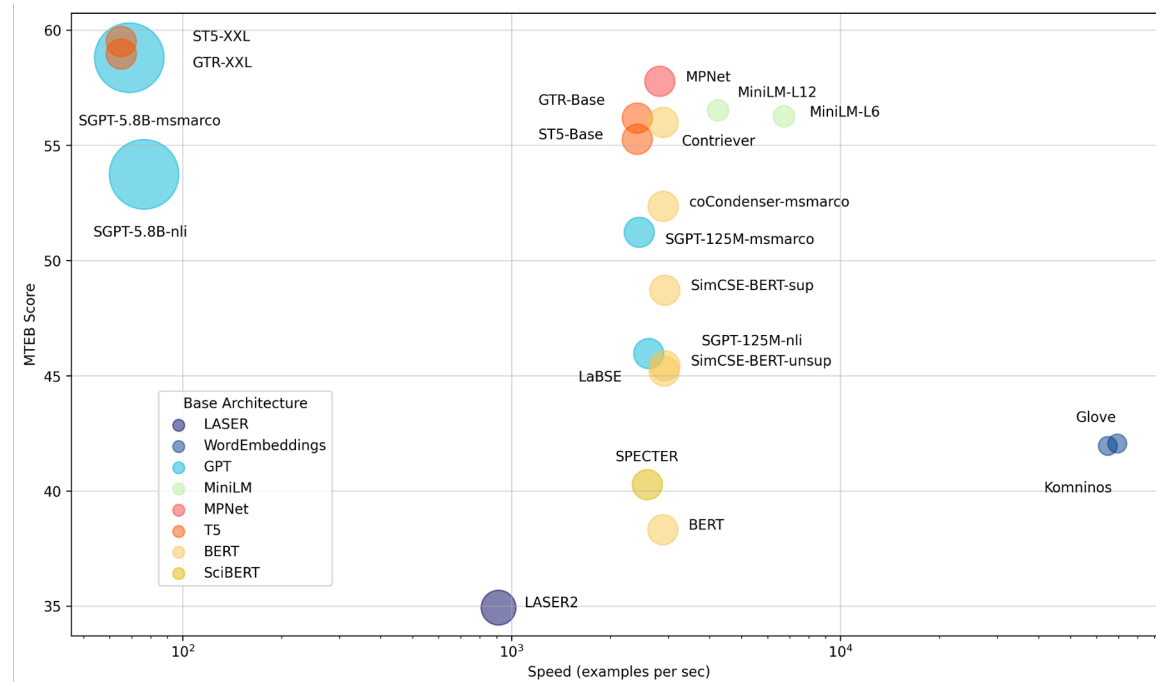
How to Train Such Model?

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



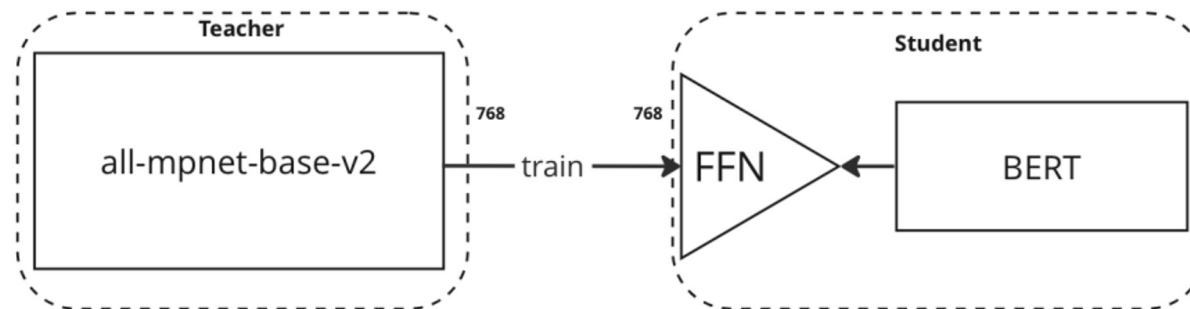
How to Train Such Model?

- ✓ 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

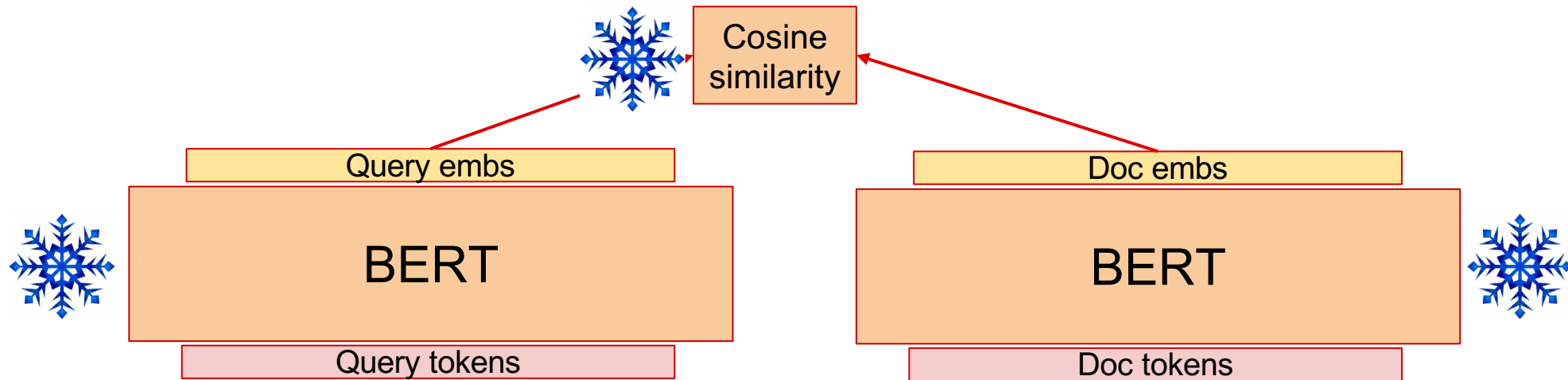


■ **Figure 4.2** The difference in multilingual distillation using PCA and one layer of Feed-Forward Neural network (FFN). The PCA trains students on embedding size of 256, meanwhile the FFN trains the student on embedding size of 768.



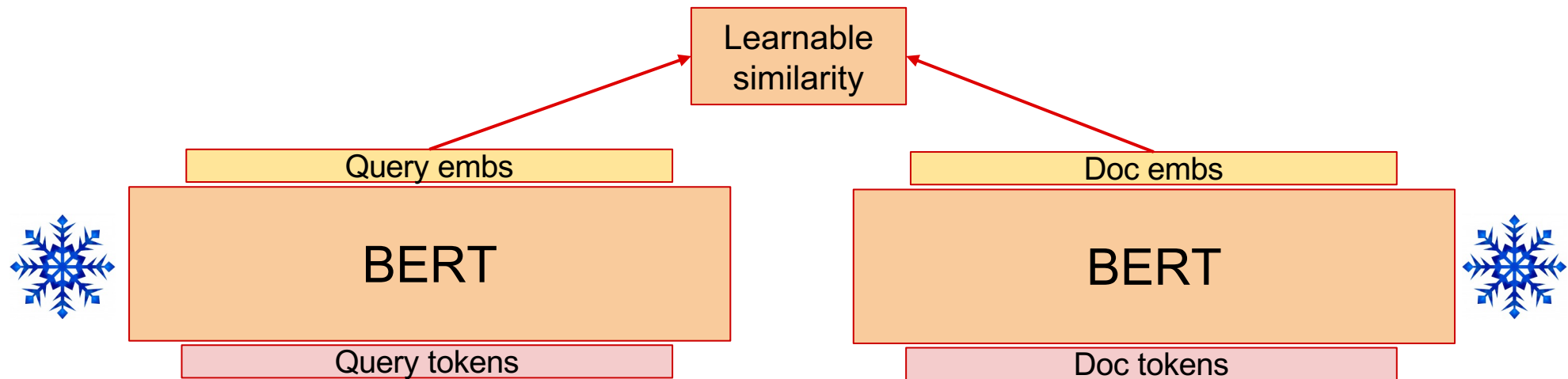
Evaluations

✓ Zero-shot



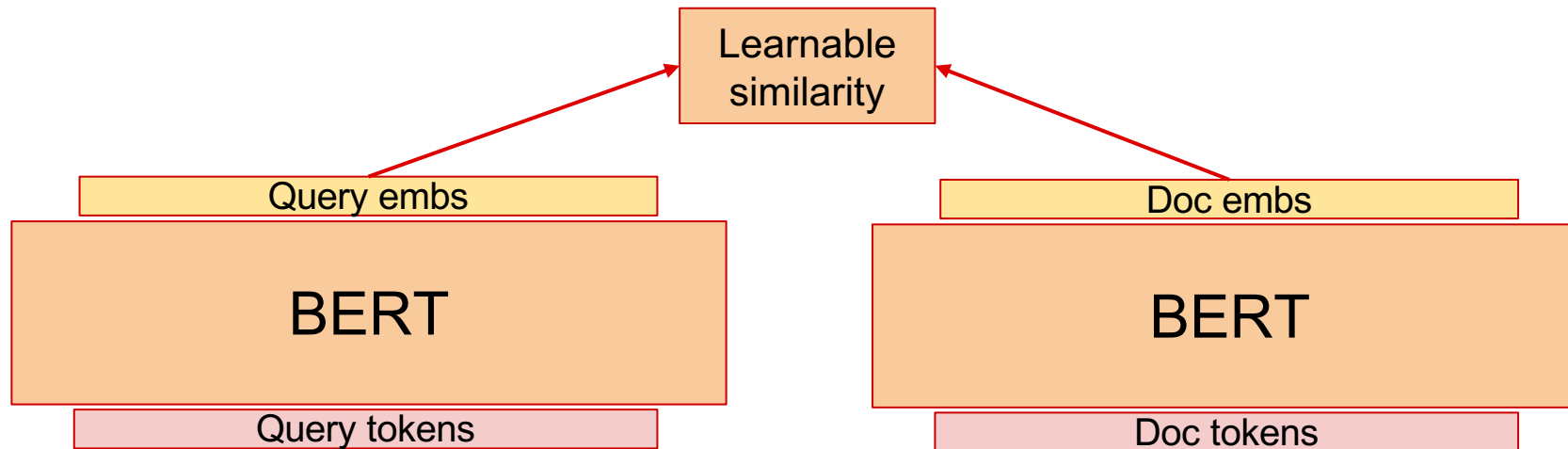
Evaluations

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



Evaluations

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



Zero Shot

SMALL MODEL	Costra	STS	DaReCzech
Small-E-Czech	63.05	48.3	36.90 \pm 0.36
RetroMAE	69.91	76.30	42.16 \pm0.36

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 \pm 0.36
RetroMAE	69.91	76.30	42.16 \pm 0.36
Dist-MPNet-CzEng	71.22	87.60	42.01 \pm 0.37
Dist-MPNet-ParaCrawl	70.42	84.25	42.33 \pm0.32

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 \pm 0.36
RetroMAE	69.91	76.30	42.16 \pm 0.36
Dist-MPNet-CzEng	71.22	87.60	42.01 \pm 0.37
Dist-MPNet-ParaCrawl	70.42	84.25	42.33 \pm0.32
OpenAiEmbeddings (ADA)	69.01	82.59	42.21 \pm 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:

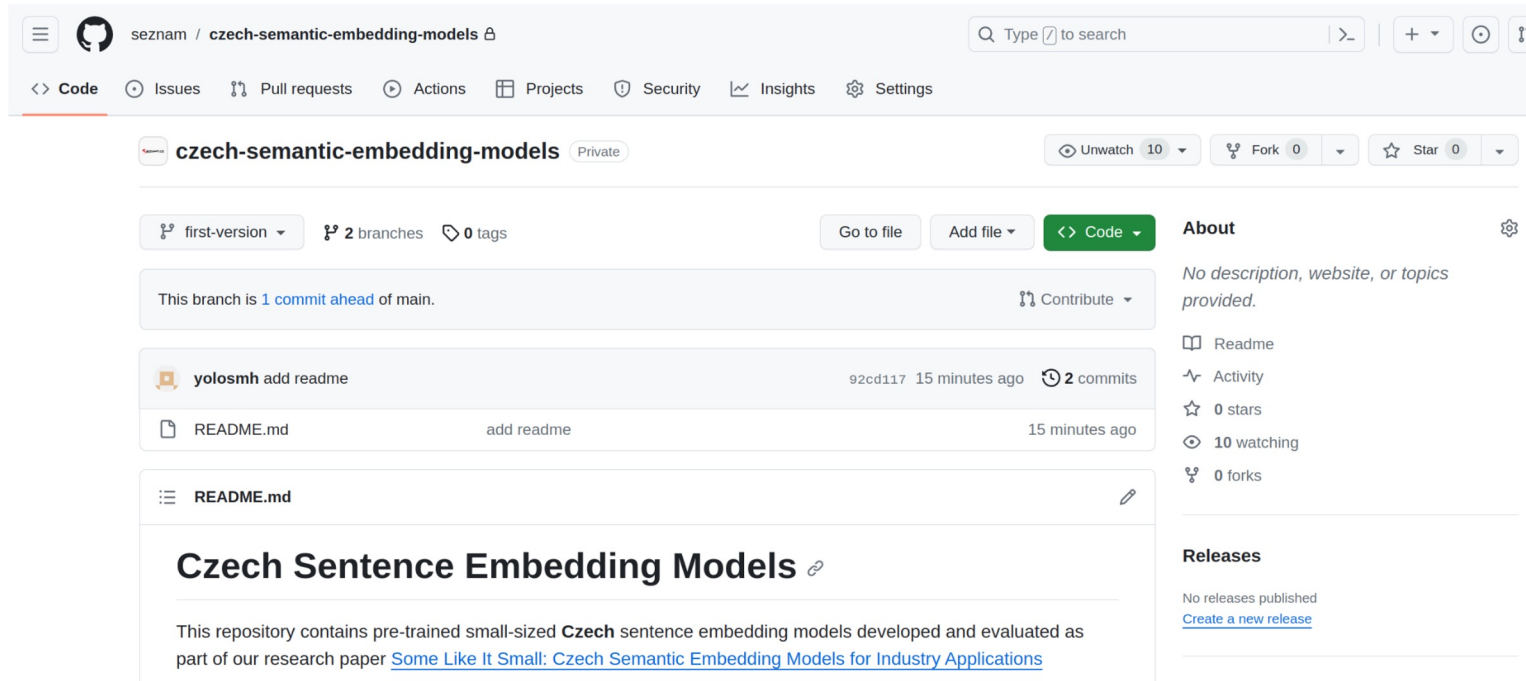
- Error rate reduced by 20 %

Image Search:

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



Models Available For You



The screenshot shows the GitHub interface for the repository 'seznam / czech-semantic-embedding-models'. The repository is private and has 10 unwatchers, 0 forks, and 0 stars. The current branch is 'first-version', which is 1 commit ahead of the main branch. A commit by 'yolosmh' titled 'add readme' was made 15 minutes ago. The repository contains a 'README.md' file. The README content is as follows:

Czech Sentence Embedding Models

This repository contains pre-trained small-sized **Czech** sentence embedding models developed and evaluated as part of our research paper [Some Like It Small: Czech Semantic Embedding Models for Industry Applications](#)

The right sidebar shows the 'About' section with no description, website, or topics provided. It also lists 'Readme', 'Activity', '0 stars', '10 watching', and '0 forks'. The 'Releases' section shows 'No releases published' and a link to 'Create a new release'.



<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



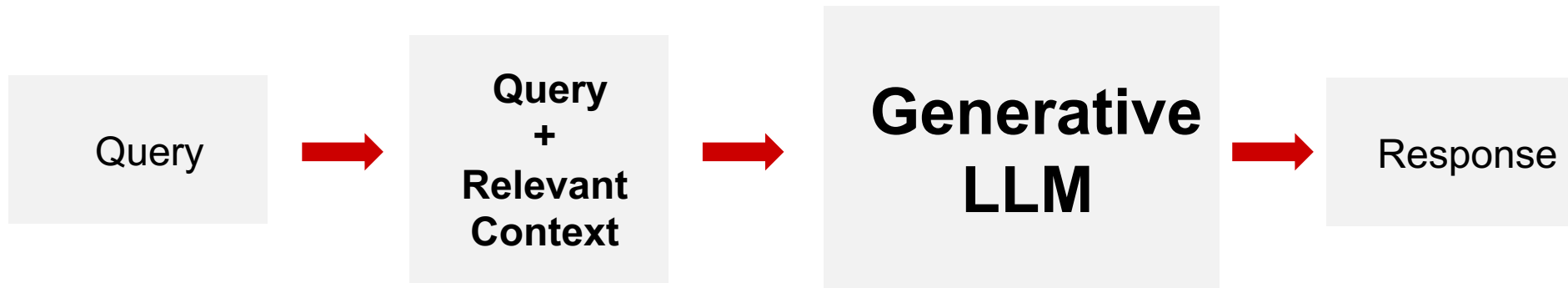
Future Plans

- ✓ 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

Copyright © 1996–2023 Seznam.cz, a. s.



Prostor pro dotazy

