Архитектура нейронный сетей: Трансформеры

Введение в курс

Содержание курса

ID	Date	Description	Sessions	Presenter
1	14.10	**Transformer model**. Motivation, original architecture and attention mechanism.	2	Maria Tikhonova
2	22.10	**Transformer-based encoders**. Masked language models based on the Transformer	2	Maria Tikhonova
3	24.10	**Classification and sequence tagging with Transformers**. Using encoders to generate feature representation for various NLU tasks.	2	Maria Tikhonova
4	29.10	**Transformer-based decoders**. Generation of text using Transformers. GPT and related decoders.	2	Victoriia Chekalina
5	31.10	**Towards ChatGPT.** Instruction tuning, einforcement Learning from Human Feedback (RLHF) and main modern LLMS	2	Maria Tikhonova
6	5.11	**Sequence to sequence tasks**. Machine translation, text detoxification, question answering, dialogue. Technical tricks for training and inference.	2	Irina Nikishina
7	7.11	**Efficient Transformers**. Pruning, Quantization, Distillation, parallelism on multiple GPUs: Varuna, Megatron pipelines	2	Victoriia Chekalina
8	12.11	**Uncertainty estimation for Transformers** and NLP tasks. Aleatoric and epistemic uncertainty. Out-of-domain detection methods.	3	Artem Vazhentsev
9	14.11	**Multilingual language models** based on the Transformer architecture.	2	Maria Tikhonova
10	19.11	**Multimodal and vision Transformers.** VQ-VAE. ViT, CLIP, Dall-E and related models. Generation with Kandinsky. Multimodal dialogue models.	2	Anton Razzhigaev
11	21.11	**Transformers for event sequences.** Classic time series and event sequence comparison. Peculiarities of event sequence modeling.	2	Vladislav Zhuzhel
12	26.11	**Transformers for tabular data.** Attention for tabular data. Tabular data architecture. Frameworks.	2	Maria Tikhonova
13	28.11	**RAG with Transformers**. Basic concept of Retriever-Augmented Generation. RAG-pipeline and frameworks. Introduction to AI Agents.	2	Maria Tikhonova
--	02.12	--	--	--
14	05.12	**Deadline** for all homework assignments. Final quiz.	2	Irina Nikishina
15	12.12	**End of the course**. Annoucement of the final scores.	4	Andrey Sakhovskiy, Elisey Rykov
		Assignments:		
		1) Knowledge graph question answering	-	Andrey Sakhovskiy
		2) Multilingual text detoxification	-	Elisey Rykov

Команда курса

Ведение лекций и семинаров

- **Мария Тихонова** (Сбер Девайсы, ВШЭ)
- **Виктория Чекалина** (AIRI, ex Сколтех)
- **Ирина Никишина** (Университет Гамбурга, ex Сколтех)
- **Антон Разжигаев** (Сколтех, AIRI)
- **Артем Важенцев** (Сколтех, AIRI)
- **Владислав Жужель** (Сколтех)
- **Андрей Саховский** (Сколтех, Сбер ИИ)
- **Елисей Рыков** (Сколтех, Т-банк ИИ)

Координация курса

- **Алексей Зайцев** (Сколтех)
- **Александр Панченко** (Сколтех, AIRI)

Формат курса

- 12 занятий по вторникам и четвергам, 18:00 - 20:20
  - Перерыв 10 мин между лекцией (1ч) и семинаром (1ч)
  - 10-15 мин на вопросы после каждой сессии
  - Читаем **на русском**, но слайды преимущественно **на английском**
    - как и большинство литературы в области NLP, Deep Learning
    - ArXiv, ACL Anthology, ICLR, NeurIPS, … – важные международные источники исследовательских статей
    - AIST, AINL, Dialogue – преимущественно русскоязычное сообщество, однако тоже издаются на английском
    - Блоги (за исключением Хабра и некоторых телеграмм сообществ)
    - => для продвижения в области важно освоить чтение в источниках на английском!

GET READY TO LEARN (READ) IN ENGLISH :)

Почему рабочий язык – английский

- Подавляющая часть существующих материалов – на английском
- Терминология – на английском
- Статьи – на английском
- Основная терминология – транслитерация с английского (промптинг, трансформер, файнтьюнинг)

Мы обязательно будем помогать с терминологией и переводом, обязательно спрашивайте, если что-то непонятно!

Формат курса

Все задания – самостоятельные

- Вопросы по лекциям (оцениваются автоматически)
- Задания по семинарам (на дополнительные эксперименты к проведенному занятию)
- Две домашние работы (на решение комплексных задач)
- Финальный тест

Оценка = 40 + 40 + 20 + ??? = ДЗ1 + ДЗ2 + Тест + Бонус

Бонус = 1 балл за вопросы по лекции + 2 балла за задание по семинару

100 баллов – наивысшая оценка

Рекомендуемая литература

- Dan Jurafsky and James H. Martin (2021). Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

- Denis Rothman (2021). Transformers for Natural Language Processing. https://www.packtpub.com/product/transformers-for-natural-language-processing/9781800565791

- The Hugging Face NLP course.
  - https://huggingface.co/course
  - https://huggingface.co/learn/nlp-course/chapter1/1

Домашние задания

Домашние задания - мини проекты

Ответы на вопросы с использованием графа знаний и LLM
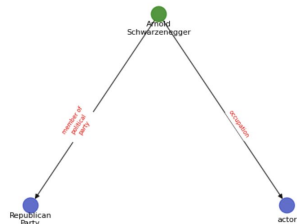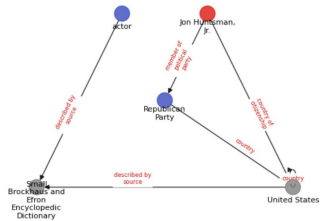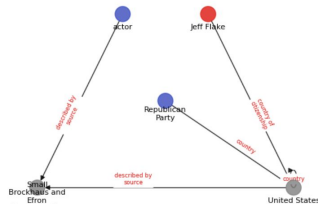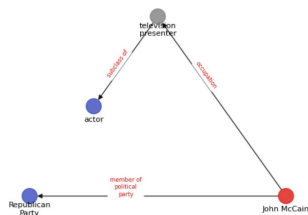
Мультиязычная детоксификация текста

> You idiot, stop talking of what you have no idea about!

Trainable detoxifier

> I would suggest learning more about the subject to get a more productive conversation.

Детоксификация текста: нужно перефразировать из грубого в нейтральный и сохранить исходных смысл

	toxic sentence	detoxified sentence
	из за таких [____] мы и страдаем	Из-за таких людей мы и страдаем
translation:	*We suffer from such [____]*	*We suffer from such people*
	[____] знает кто кум, но девушка красивая 👍	неизвестно кто кум, но девушка красивая
translation:	*[____] knows who the godfather is, but the girl is beautiful* 👍	*it is unknown who the godfather is, but the girl is beautiful*
	порядок бы [____] навёл !	Порядок бы навел
translation:	*Put these [____] things in order*	*Put the things in order*

Детоксификация текста: нужно перефразировать из грубого в нейтральный и сохранить исходных смысл

"Multilingual Text Detoxification"

- Лидерборд: https://codalab.lisn.upsaclay.fr/competitions/18243

- Код: https://github.com/pan-webis-de/pan-code/tree/master/clef24/text-detoxification

- Информация о деталях задаче:
  https://pan.webis.de/clef24/pan24-web/text-detoxification.html

Ответы на вопросы с использованием графа знаний и LLM

"Knowledge Graph Question Answering"

- Лидерборд: https://codalab.lisn.upsaclay.fr/competitions/18214#

- Код: https://github.com/uhh-lt/TextGraphs17-shared-task/

- Информация о деталях задаче:
  https://sites.google.com/view/textgraphs2024/home/shared-task

Ответы на вопросы с использованием графа знаний и LLM
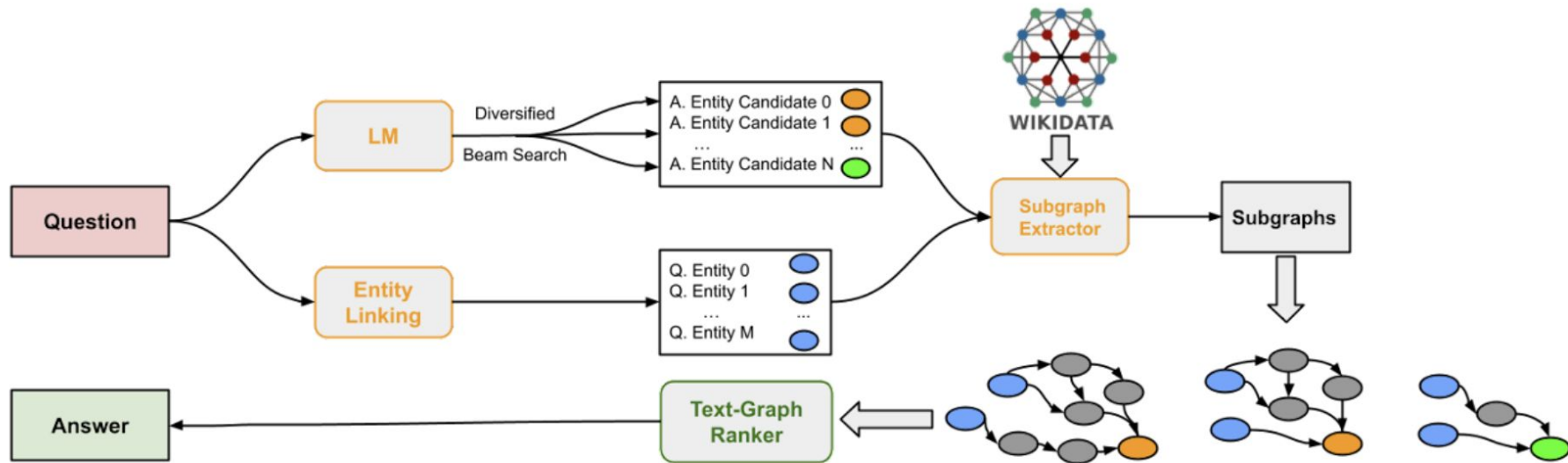
Ответы на вопросы с использованием графа знаний и LLM

question	questionEntity	answerEntity	groundTruthAnswerEntity
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Arnold Schwarzenegger	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Chuck Grassley	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Jeff Flake	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	John McCain	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Jon Huntsman, Jr.	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Lindsey Graham	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Mitch McConnell	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Orrin Hatch	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Rand Paul presidential campa	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Rick Santorum	Arnold Schwarzenegger
Who was formerly an actor and now a Republican senator?	actor, Republican Party	Roy Blunt	Arnold Schwarzenegger

Система оценивания

Технический отчет		Код		Результаты		Сумма баллов
Методология	Анализ результатов	Читаемость	Воспроизводимость	Преодоление базового решения	Top-1 +10 баллов Top-20% +5 баллов	100% + бонус
5	5	5	5	0 / 5	0 / 5 / 10	25 (max 35)

Технический отчет

[Colab link](#)

Технический отчет

Код

▾ 3. Code

Enter here all code used to produce your results submitted to Codalab. Add some comments and subsections to navigate though your solution.

In this part you are expected to develop yourself a solution of the task and provide a reproducible code:

- *Using Python 3;*
- *Contains code for installation of all dependencies;*
- *Contains code for downloading of all the datasets used;*
- *Contains the code for reproducing your results (in other words, if a tester downloads your notebook she should be able to run cell-by-cell the code and obtain your experimental results as described in the methodology section).*

As a result, you code will be graded according to these criteria:

- ***Readability**: your code should be well-structured preferably with indicated parts of your approach (Preprocessing, Model training, Evaluation, etc.).**
- ***Reproducibility**: your code should be reproduced without any mistakes with "Run all" mode (obtaining experimental part).**

▸ 3.1 Requirements

[ ] ↳ *1 cell hidden*

▸ 3.2 Download the data

[ ] ↳ *2 cells hidden*
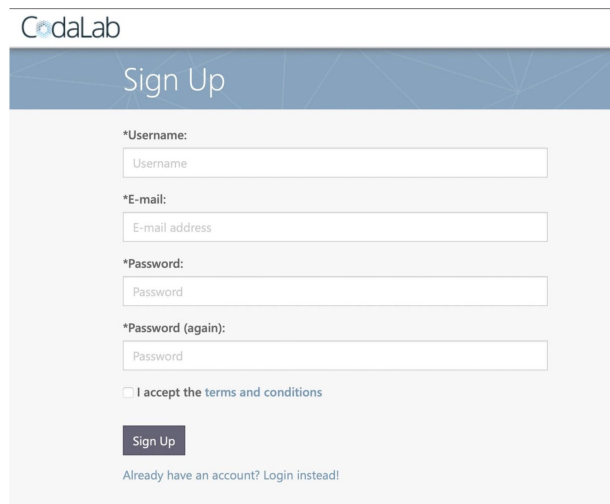
▸ 3.3 Preprocessing

[ ] ↳ *1 cell hidden*

▸ 3.4 My method of text processing

▶ ↳ *1 cell hidden*

CodaLab (регистрация)

Knowledge Graph Question Answering:
Multilingual Text Detoxification:

CodaLab (необходимо указать имя команды)

Отправка решения

Лидерборд
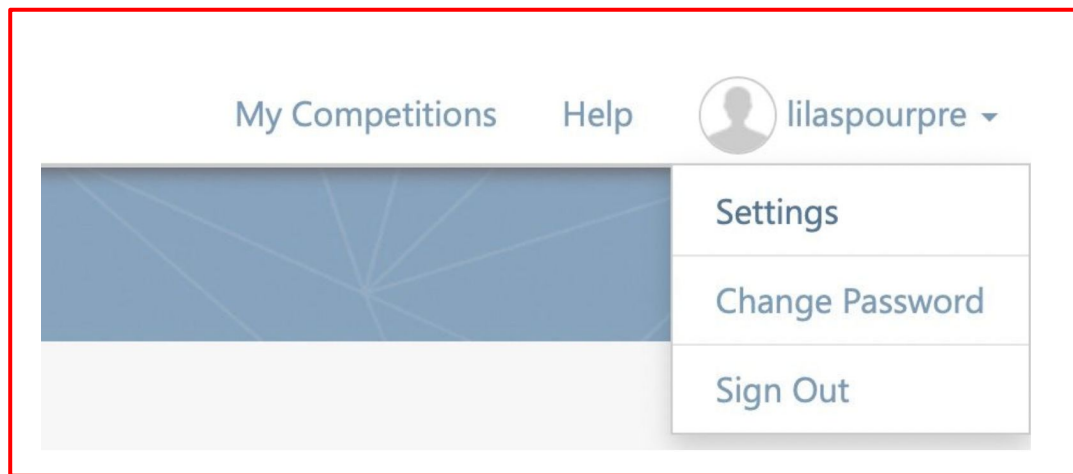
Learn the Details   Phases   Participate   **Results**

Development   Final test   Post-evaluation

Phase description
!!! Attention !!! You should submit a zip archive which contains a *.txt file with your detoxified sentences.

Max submissions per day: 999

Max submissions total: 100

Download CSV   Download all submissions on leaderboard

Results

#	User	Entries	Date of Last Entry	Team Name	Style transfer accuracy ▲	Meaning preservation ▲	Fluency ▲	Joint score ▲	ChrF1 ▲
1	gleb.shnshn	3	02/04/22		0.97 (3)	0.93 (1)	0.95 (4)	0.86 (1)	0.53 (13)
2	orzhan	24	01/25/22		0.98 (2)	0.83 (6)	0.95 (2)	0.78 (2)	0.55 (10)
3	Phoenix120	24	01/23/22	Mindful Squirrel	0.91 (7)	0.90 (2)	0.95 (1)	0.78 (3)	0.55 (8)
4	Rexhaif	1	01/03/22	FRC CSC RAS	0.93 (5)	0.83 (7)	0.95 (3)	0.73 (4)	0.57 (4)
5	aorews	1	02/08/22		0.94 (4)	0.79 (13)	0.85 (9)	0.66 (5)	0.52 (15)