Олимпиада НИУ ВШЭ для студентов и выпускников – 2019 г.

Направление: «Фундаментальная и прикладная лингвистика»

Профиль:
«Компьютерная лингвистика и цифровые методы в гуманитарных науках» КОД – 310

Время выполнения задания – 180 мин., язык - русский.

I. ОБЩАЯ ЧАСТЬ.

Задание 1. Прочтите отрывок статьи "Dairy farming, solar panels, and diagnosing Parkinson's disease: what can you do with deep learning?" (Rachel Thomas) (https://www.fast.ai/2019/02/21/dl-projects/).

Составьте краткое резюме этого отрывка на русском языке (объемом в один абзац, но не более 1000 знаков).

а) Каков основной тезис?

б) Какие аргументы за или против выдвигаемого тезиса приводит автор?

в) Выскажите ваши собственные соображения по проблемам, затронутым в отрывке, и их решениям.

г) Какие способы применения AI для автоматической обработки текста вы можете назвать?

«Many people incorrectly assume that AI is only for an elite few– a handful of Silicon Valley computer science prodigies with monthly budgets larger than most people's lifetime earnings, turning out abstruse academic papers. This couldn't be more wrong. Deep learning (a powerful type of AI) can, and is, being used by people with varied backgrounds all over the world. A small taste of that variety can be found in the stories shared here: a Canadian dairy farmer trying to identify udder infections in his goats, a Kenyan microbiologist seeking more efficiency in the lab, a former accountant expanding use of solar power in Australia, a 73-year old embarking on a second career, a son of refugees who works in cybersecurity, and a researcher using genomics to improve cancer treatment. Hopefully this may inspire you to apply deep learning to a problem of your own!

Building Tools for Microbiologists in Kenya

Benson Nyabuti Mainye trained as a microbiologist in his home country of Kenya. He noticed that lab scientists can spend up to 5 hours studying a slide through a microscope to try to identify what cell types were in it, and he wanted a faster alternative. Benson created an immune cell classifier to distinguish various immune cells (eosinophils, basophils, monocytes, and lymphocytes) within an image of a blood smear. This fall, he traveled to San Francisco to attend part of the fast.ai course in person at the USF Data Institute (a new session starts next month), and where another fast.ai classmate, Charlie Harrington, helped him deploy the immune cell classifier. Since malaria is one of the top 10 causes of death in Kenya, Benson is currently working with fellow Kenyan and fast.ai alum Gerald Muriuki on a classifier to distinguish different types of mosquitoes to isolate particular types that carry the Plasmodium species (the parasite which causes malaria).

Dairy Goat Farming

Cory Spencer is a dairy goat farmer on bucolic Vancouver Island, and together with his wife owns The Happy Goat Cheese Company. When one of his goats came down with mastitis (an udder infection), Cory was unable to detect it until after the goat had suffered permanent damage. Estimates suggest that mastitis costs the dairy industry billions of dollars each year. By combining a special camera that detects heat (temperatures are higher near an infection) together with deep

learning, Cory developed a tool to identify infections far earlier (at a subclinical level) and for one-tenth the cost of existing methods. Next up: Cory is currently building a 3D model to track specific parts of udders in real time, towards the goal of creating an automatic goat milking robot, since as Cory says, "The cow guys already have the fancy robotic tech, but the goat folk are always neglected."

From Accountant to Deep Learning Practitioner working on Solar Energy

Sarada Lee was a former accountant looking to transition careers when she began a machine learning meetup in her living room in Perth, Australia, as a way to study the topic. That informal group in Sarada's living room has now grown into the Perth Machine Learning Meetup, which has over 1,400 members and hosts 6 events per month. Sarada traveled to San Francisco to take the Practical Deep Learning for Coders and Cutting Edge Deep Learning for Coders courses in person at the USF Data Institute, and shared what she learned when she returned back to Perth. Sarada recently won a 5-week long hackathon on the topics of solar panel identification and installation size prediction from aerial images, using U-nets. As a result, she and her team have been pre-qualified to supply data science services to a major utility company, which is working on solar panel adoption for an area the size of UK with over 1.5 million users. Other applications they are working on include electricity network capacity planning, predicting reverse energy flow and safety implications, and monitoring the rapid adoption of solar.

Sarada and the Perth Machine Learning Meetup are continuing their deep learning outreach efforts. Last month, a team led by Lauren Amos created an interactive creative display at the Fringe World Festival to make deep learning more accessible to the general public. This was a comprehensive team effort, and the display included:

- *artistic panels design based on style transfer*
- *GRU/RNN generated poems*
- *Implemented BERT to generate poems or short books*
- *Applied speech-to-text and text-to-speech APIs to interact with a poetry-generating robot*

Festival attendees were able to enjoy the elegant calligraphy of machine generated poems, read chapters of machine-generated books, and even request a robot to generate poems given a short seed sentence. Over 4,000 poems were generated during the course of the 2-week festival!».

II. СПЕЦИАЛЬНАЯ ЧАСТЬ

Выберите и выполните <u>только один</u> из блоков заданий специальной части в соответствии с выбранной вами программой магистерской подготовки.

Направление: "Компьютерная лингвистика"

Задание 2. Решите задачу.

Буквы Ш, Л, Ь, И, Е, Н написаны на отдельных карточках. Акакий Акакиевич берет карточки в случайном порядке и прикладывает одну к другой. Какова вероятность, что Акакий Акакиевич соберет из них слово «ШИНЕЛЬ»?

Задание 3.

Вам нужно создать систему “Стихосложатор”, которая отвечает пользователю в рифму на любую введенную фразу на русском языке.

Например:

Пользователь: “хоть и заглядывал я встарь”

Система: “в академический словарь”

Пользователь: “прилетели инопланетяне”

Система: “в современном любовном романе”

Ответ системы должен быть осмысленным, грамматически правильным и синтаксически согласованным с введенной фразой. Также он должен соответствовать метрической структуре введенной фразы и рифмоваться с ней. Примеры недопустимых ответов:

Пользователь: “хоть и заглядывал я встарь”

Система: “толковый словарь”

Пользователь: “прилетели инопланетяне”

Система: “в современный любовный рассказе”

Ваша задача как компьютерного лингвиста разработать архитектуру и методы тестирования такой системы.

1. Предложите представление метра в виде любой формальной записи, а также опишите условные обозначения так, чтобы оценивающий Вашу работу мог их проинтерпретировать.
2. а) Если использовать правиловый подход для решения данной задачи, то какие правила могли бы войти в Вашу систему? Приведите два-три примера. Какая предварительная обработка текста необходима Вашей системе? Какие морфологические характеристики словоформ могут понадобиться для порождения правильного ответа (если вы считаете, что морфологическая информация не понадобится, обоснуйте)? Что должен делать морфологический модуль в данной задаче при правиловом подходе? Если Вы считаете, что Вам нужен синтаксический модуль, то в каком формализме Вы будете представлять синтаксическую структуру, какая синтаксическая информация понадобится Вам в правилах? Ответ обоснуйте.

   б) Представьте, что вы решаете эту задачу методами машинного обучения. Что именно будет объектом в Вашей задаче? Приведите примеры признаков (features) объектов для обучения, необходимые для решения данной задачи. Нужны ли синтаксические признаки? Если да, то какие? Какие ответы будет выдавать Ваша система? Ответ обоснуйте.
3. По каким параметрам можно протестировать работу системы? Какой могла бы быть система рейтингов (штрафов, баллов и т.п.) для разных лингвистических функций? Приведите примеры. Как получить и интерпретировать результирующую оценку качества лингвистической системы в целом?

Направление: “Цифровые методы в гуманитарных науках”

Задание. 2

Для решения задачи не требуется знакомства с старотатарским языком, все необходимые лингвистические представления можно перенести из русского языка. Определите размер этого двустишия и объясните свой ответ:

Ässälam wä ässälam wä ässälam;
I ƺanyj, bäɣdä sälam bulyr qälam

Буквы ŋ, ƺ и ɣ означают согласные звуки.

В тюркском варианте восточной системы стихосложения аруз размер, в котором созданы такие поэтические строки на старотатарском языке, называется *хазадж-и мусаддас-и махзуф* (написание несколько упрощено в угоду метрике):

Anyŋ kem al äŋindä mäŋ jaratty, Bujy berlä sačyny täŋ jaratty Xäkimnärdän qalan süz dörr wä ƺäwhär; Ґaqylly kemsä alyr any ezbar. Babaxan digänul ber šäh barirde

Размер этих строк — *рамал-и мусаддас-и махзуф:*

Šähidirür, ike küze jäširür. Jad itärlär — kem belä? - räxmät ilä

Размер этих строк — *рамал-и мусамман-и махзуф*:

Käšt itep gäzdem bu tatar ileneŋ jaxšylaryn. Kürde küzem, döšde küŋlem ul bäder surätenä Näq Qazan aertynda bardyr ber awyl — Qyrlaj dilär

Размер этих строк — *хазадж-и мусамман-и салим*:

Güzäl šuridä bylbyl da fävan äjlär, šahym Tahir; Ґäziz ƺanyn fida äjlär, kürep ul Zöhrä dildati Xodāɣa küp xämid itkän xämidämdin ƺöda buldym Menä kič. Zur awyl östendä čyqty nurly aj qalqyp

Задание 3.

Представьте, что вас заинтересовал феномен русскоязычной любительской литературы, т.е. художественные тексты, написанные непрофессиональными авторами, не публикуемые издательствами и не приносящие их создателям никаких денег. Каким образом можно исследовать литературу такого рода количественными методами?

Опишите:

1. где и каким образом вы бы предложили собрать материал для вашего исследования; какие компьютерные технологии могут быть при этом применены?
2. какие типы разметки можно было бы предложить для собранного вами материала? проявите максимум фантазии и предложите как можно больше уровней разметки
3. какого рода метаинформация о ваших объектах исследования вам может понадобиться?
4. предложите не менее трех сценариев количественных исследований на базе вашего материала. В каждом случае пропишите цель исследования и необходимые шаги.