

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО

**Директор физтех-школы
прикладной математики и
информатики
А.М. Райгородский**

	Рабочая программа дисциплины (модуля)
по дисциплине:	Машинный перевод
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Физтех-школа Прикладной Математики и Информатики кафедра анализа данных
курс:	4
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 1

Программу составил: В.В. Кантор

Программа обсуждена на заседании кафедры анализа данных 06.03.2020

Аннотация

Курс посвящён рассмотрению современных подходов к решению задачи машинного перевода. В первую очередь будут рассматриваться нейросетевые методы на основе архитектур энкодер-декодер, а также трансформеров. В курсе будут обсуждаться в том числе такие проблемы, как работа с low resource языками и использование контекста.

1. Цели и задачи

Цель дисциплины

ознакомление студентов с основными принципами правильного и статистического машинного перевода.

Задачи дисциплины

- освоение студентами базовых знаний (понятий, концепций, методов и моделей) в области машинного перевода;
- приобретение теоретических знаний и практических умений и навыков в области создания алгоритмов автоматического перевода;
- оказание консультаций и помощи студентам в построении собственных алгоритмов.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
	УК-2.2 Проектирует решение конкретной задачи проекта, выбирая оптимальный способ ее решения, исходя из действующих правовых норм и имеющихся ресурсов и ограничений
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценить качество разработанной модели
	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности

новые научные результаты	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные понятия, законы машинного перевода;
- современные методы фразового и нейросетевого машинного перевода.

уметь:

- понять поставленную задачу;
- написать собственную систему машинного перевода, основанную как на правилном, так и на статистическом подходе;
- оценивать качество систем машинного перевода;
- применять различные технологии автоматической обработки текстов, включая языковые модели, POS-тэггинг, синтаксические анализаторы к задаче машинного перевода;
- строить и обучать нейронные сети, использовать вложения для решения задач машинного перевода;
- самостоятельно находить способы выполнения поставленных задач, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов.

владеть:

- навыками освоения большого объема информации и решения задач (в том числе, сложных);
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения теоретических задач лингвистики;
- методами автоматического морфологического и синтаксического анализа и синтеза.

4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Вводная лекция. История вопроса. Базовые принципы фразового перевода.	3	3		3
2	Лингвистика и машинный перевод.	3	3		3
3	Синтаксис и машинный перевод.	3	3		3
4	Глобальные свойства синтаксической структуры.	3	3		3
5	Словарь в машинном переводе.	3	3		3

6	Основные принципы статистического машинного перевода.	3	3		3
7	Выравнивание.	3	3		3
8	End-to-end фразовый перевод.	3	3		3
9	Word embeddings.	3	3		3
10	Последние достижения статистического машинного перевода.	2	2		2
11	Encoder-decoder модели.	1	1		1
Итого часов		30	30		30
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 8 (Весенний)

1. Вводная лекция. История вопроса. Базовые принципы фразового перевода.

Простейший пословный декодер (при наличии словаря). Beam search. Языковая модель. Выравнивание. Как получить словарь из интернета за 3 легких шага: параллельные документы, параллельные предложения, параллельные слова. Фразы. Перестановки. BLEU.

2. Лингвистика и машинный перевод.

Уровни и подуровни представления единиц текста. Анализ и синтез текста. Неоднозначность языковых единиц как ключевая проблема машинного перевода. Автоматический морфологический анализ и синтез. Морфологическая структура. Морфологические категории и их значения.

3. Синтаксис и машинный перевод.

Автоматический синтаксический анализ и синтез. Основные типы синтаксического представления предложения. Дерево составляющих и дерево зависимостей. Синтаксические отношения. Синтаксические признаки.

Семантический анализ и синтез. Глубокая семантика. Дескрипторы и концепты. Онтологическая семантика. Логика здравого смысла.

4. Глобальные свойства синтаксической структуры.

Синтаксические признаки. Предикатные слова, валентности и актанты. Лексические функции.

5. Словарь в машинном переводе.

Грамматика и словарь в машинном переводе. Толково-комбинаторный словарь. Трансфер. Лингвистическая семантика. Онтологическая семантика. Интерлингва. Интерактивность при машинном переводе. Разбор конкретной правилковой системы МП (ЭТАП-3).

6. Основные принципы статистического машинного перевода.

Почему машинный перевод – это сложно? Построение системы машинного перевода по данным. Важнейшие прорывы в истории статистического перевода. Оценка систем машинного перевода.

7. Выравнивание.

Оценка максимального правдоподобия. EM-алгоритм. Модели выравнивания IBM.

8. End-to-end фразовый перевод.

Перестановки. N-граммные языковые модели. Фразовый перевод. Оптимизация компонент.

9. Word embeddings.

Нейросетевые языковые модели. Пространства вложений. Использование вложений в задачах автоматической обработки текстов.

10. Последние достижения статистического машинного перевода.

Масштабирование алгоритмов для работы с (очень) большими данными. Использование данных на одном языке. Обучение с подкреплением. Гибридные символьные/нейронные модели.

11. Encoder-decoder модели.

Архитектуры рекуррентных слоёв нейронных сетей: RNN, LSTM, GRU и т. д. Sequence-to-sequence модели в машинном переводе. Attention модели.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

учебная аудитория, оснащенная медиапроектором и экраном.

6. Перечень рекомендуемой литературы

Основная литература

1. Введение в методы машинного обучения с подкреплением, учебное пособие /А. И. Панов; Министерство науки и высшего образования Российской Федерации ; Московский физико-технический институт (национальный исследовательский университет). Москва, МФТИ, 2019

Дополнительная литература

1. Дисперсионный анализ и синтез планов на ЭВМ [Текст]/Е. В. Маркова [и др.] , -М., Наука, 1982

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

Не используются

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

1. <https://arxiv.org/abs/1609.08144>
2. Open NMT, An open-source neural machine translation system: <http://opennmt.net/about/>

9. Методические указания для обучающихся по освоению дисциплины (модуля)

1. Рекомендуется успешно сдавать самостоятельную работы, так как это упрощает итоговую аттестацию по предмету.

2. Для подготовки к итоговой аттестации по предмету лучше всего пользоваться материалами лекций.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Информатика и вычислительная техника

профиль подготовки: Физтех-школа Прикладной Математики и Информатики
кафедра анализа данных

курс: 4

квалификация: бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Разработчик: В.В. Кантор

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-2 Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1 Формулирует совокупность взаимосвязанных задач в рамках поставленной цели работы, обеспечивающих ее достижение. Определяет ожидаемые результаты решения поставленных задач
	УК-2.2 Проектирует решение конкретной задачи проекта, выбирая оптимальный способ ее решения, исходя из действующих правовых норм и имеющихся ресурсов и ограничений
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
	ОПК-2.3 Знает основные требования информационной безопасности
ОПК-3 Способен составлять и оформлять научные и (или) технические (технологические, инновационные) отчеты (публикации, проекты)	ОПК-3.1 Знает основные правила оформления научных публикаций и научно-технической документации, в том числе с использованием прикладного программного обеспечения
	ОПК-3.2 Владеет на практике методологией составления научно-технических отчетов (проектов)
	ОПК-3.3 Владеет методами визуального и графического представления результатов научной (научно-технической, инновационной технологической) деятельности в виде отчетов, научных публикаций
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценить качество разработанной модели
	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты
ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию	ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации
	ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива
	ПК-2.3 Способен проводить апробацию результатов научно-исследовательской работы посредством публикации научных статей и участия в конференциях

2. Показатели оценивания компетенций

В результате изучения дисциплины «Машинный перевод» обучающийся должен:

знать:

- основные понятия, законы машинного перевода;
- современные методы фразового и нейросетевого машинного перевода.

уметь:

- понять поставленную задачу;
- написать собственную систему машинного перевода, основанную как на правилном, так и на статистическом подходе;
- оценивать качество систем машинного перевода;
- применять различные технологии автоматической обработки текстов, включая языковые модели, POS-тэггинг, синтаксические анализаторы к задаче машинного перевода;
- строить и обучать нейронные сети, использовать вложения для решения задач машинного перевода;
- самостоятельно находить способы выполнения поставленных задач, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов.

владеть:

- навыками освоения большого объема информации и решения задач (в том числе, сложных);
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения теоретических задач лингвистики;
- методами автоматического морфологического и синтаксического анализа и синтеза.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

1. Как получить словарь из интернета за 3 легких шага: параллельные документы, параллельные предложения, параллельные слова.
2. Автоматический морфологический анализ и синтез.
3. Автоматический синтаксический анализ и синтез.
4. Дерево составляющих и дерево зависимостей.
5. Семантический анализ и синтез.
5. Словарь в машинном переводе.
6. Толково-комбинаторный словарь.
7. Оценка систем машинного перевода.
8. Использование вложений в задачах автоматической обработки текстов.
9. Использование данных на одном языке.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Как получить словарь из интернета: параллельные документы, параллельные предложения, параллельные слова.
2. Оценка систем машинного перевода.
3. Морфологические категории и их значения.
4. Основные типы синтаксического представления предложения. Дерево составляющих и дерево зависимостей.
5. Выравнивание с помощью оценок максимального правдоподобия. EM-алгоритм.
6. Модели выравнивания IBM.
7. End-to-end фразовый перевод.
8. Нейросетевые языковые модели.
9. Sequence-to-sequence модели в машинном переводе.
10. Использование обучения с подкреплением в машинном переводе.

Критерии оценивания

- оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений

- оценка «отлично (9)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений
- оценка «отлично (8)» выставляется студенту, показавшему всесторонние систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и правильное обоснование принятых решений
- оценка «хорошо (7)» выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (6)» выставляется студенту, если он знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (5)» выставляется студенту, если он знает материал, и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач
- оценка «неудовлетворительно (1)» выставляется студенту, который не знает формулировок основных понятий дисциплины.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины.