**МИНОБРНАУКИ РОССИИ**

**Федеральное государственное бюджетное**

**образовательное учреждение высшего образования**

**«Тульский государственный университет»**

Подразделение: кафедра «Вычислительная техника»

**ОТЧЕТ**

|  |  |
| --- | --- |
| Вид практики | НИР (получение первичных навыков НИР) |
| Курс | 3 курс |
| Направление подготовки  /специальность | 09.03.01 Информатика и вычислительная техника |
| Ф.И.О. обучающегося | Шайхаттаров Дамир Владимирович |
| Место прохождения практики | ФГБОУ ВО «Тульский Государственный университет», каф. ВТ |
| Период прохождения практики | С 21 сентября по 17 октября 2020 г. |

Руководитель практики от

профильной организации (при наличии)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*(Ф.И.О., должность) (подпись)*

М. П.

Руководитель практики от подразделения

Неелова Н.В., ст. преподаватель каф. ВТ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*(Ф.И.О., должность) (подпись)*

г. Тула  
 2020 г.

РЕФЕРАТ

В данной НИР я буду рассматривать несколько видов машинного перевода: SMT, RBMT, HMT и наконец NMT. Каждый вид имеет свои особенности и алгоритмы перевода текста. NMT из них является самой новой и многообещающей технологией для машинного перевода, но является ли она универсальной и есть ли области перевода, где другие виды MT будут более актуальны. Эти вопросы мы разберём в данной научной работе.

Цель работы – получение навыков выполнения НИР по проектированию систем машинного перевода, работы с глобальными информационными системами для поиска и обработки научно-технической информации и обобщения и ведения научной дискуссии по проблемным вопросам программирования, проектирования и автоматизации.

В процессе работы проводился сравнительный анализ проблем проектирования систем машинного перевода различных типов MT и решений проблемы выбора вида машинного перевода исходя из особенностей каждого вида MT.

В результате исследования были выделены плюсы и минусы каждого вида машинного перевода и были объяснены какие типы MT лучше всего выбрать для определённых проектов.

СОДЕРЖАНИЕ

[РЕФЕРАТ 2](#_Toc52579591)

[СОДЕРЖАНИЕ 3](#_Toc52579592)

[ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ 4](#_Toc52579593)

[ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ 5](#_Toc52579594)

[ВВЕДЕНИЕ 6](#_Toc52579595)

[АНАЛИЗ ПРОБЛЕМ И ПРЕИМУЩЕСТВ РАЗЛИЧНЫХ ТИПОВ MT 7](#_Toc52579596)

[ЗАКЛЮЧЕНИЕ 16](#_Toc52579597)

[СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ 17](#_Toc52579598)

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящем отчёте о НИР применяют следующие термины с соответствующими определениями:

Искусственная нейронная сеть (ИНС) - [математическая модель](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D1%8C), а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей — сетей [нервных клеток](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD) живого организма.

Система [нейронного машинного перевода](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%BC%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D1%8B%D0%B9_%D0%BF%D0%B5%D1%80%D0%B5%D0%B2%D0%BE%D0%B4) - (NMT).

Персональный компьютер - компьютер, предназначенный для эксплуатации одним пользователем, то есть для личного использования.

Машинный перевод с помощью правил (RBMT) - это системы машинного перевода, основанные на лингвистической информации об исходных и целевых языках, в основном получаемой из (одноязычных, двуязычных или многоязычных) словарей и грамматик, охватывающих основные семантические, морфологические и синтаксические закономерности каждого языка соответственно.

Статистический машинный перевод (SMT) - это разновидность машинного перевода текста, основанная на сравнении больших объемов языковых пар.

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

В настоящем отчете о НИР применяют следующие сокращения и обозначения:

НИР - Научно-исследовательская работа

Нейросеть - Нейронная сеть

ПК - Персональный компьютер

ВВЕДЕНИЕ

В современном мире машинного перевода можно выделить четыре типа MT: SMT, RBMT и NMT. Google Translate, являющийся самым популярным сервисом машинного перевода на данный момент, например, использует в своём проекте нейронные сети, но чем обусловлен этот выбор? Актуальны ли альтернативы NMT? И где используются остальные виды MT я разберу далее.

Но для того, чтобы понять почему нейронные сети и остальные виды MT не могут решить проблему машинного перевода, мы должны рассмотреть какие проблемы у машинного перевода присутствуют, и как их решают различные виды MT.

АНАЛИЗ ПРОБЛЕМ И ПРЕИМУЩЕСТВ РАЗЛИЧНЫХ ТИПОВ MT

Ставится задача анализа проблем и преимуществ каждого типа машинного перевода, для проведения сравнительного анализа и дальнейшего создания алгоритма выбора MT под задачи различных проектов.

1. Анализ проблем и преимуществ SMT:
   1. Описание работы SMT:

Для начала дадим определение SMT – это тип машинного перевода, в котором перевод генерируются на основе статистических моделей, параметры которых являются производными от анализа двуязычных корпусов текста (исходного переведённого текста и целевого непереведённого текста). Из чего следует, что чтобы статический перевод был максимально достоверным требуется очень большой набора данных о ранее утвержденных переводах для образования корпуса первого типа (на самом деле при создании сервиса MT используется совокупность таких корпусов), который в дальнейшем будет преобразован в статическую модель перевода. Затем эта модель будет применена к непереведённому целевому тексту, чтобы произвести сопоставление и предложить разумный перевод. В SMT модель перевода, использует частоту фраз, появляющихся в учебном корпусе обучения (таблице переводов). В этой таблице хранится фраза и количество ее повторений во всем корпусе обучения. Чем чаще фраза повторяется в обучающем корпусе, тем более вероятно, что целевой перевод будет правильным. Каждая фраза (хранящаяся в таблице фраз) может иметь длину от одного до пяти слов. Эта таблица фраз называется моделью перевода.

SMT использует вероятностную модель, чтобы найти правильную комбинацию перевода. Этот процесс является эволюционным, поскольку корпус уточняется и корректируется после каждого прогона для устранения / корректировки любых аномалий. Чем чаще используется корпус, тем совершеннее он становится. Повышение качества корпуса - это непрерывный процесс исследований и разработок очень ценного переводческого актива.

Кроме того, в SMT строится вторичная модель, используя целевые данные перевода. Эта модель помогает определить порядок, в котором необходимо собрать фразы (из таблицы фраз), чтобы оптимизировать беглость перевода, то есть придать переведенному тексту его естественный язык. Свободное владение языком гарантирует, что дословный перевод (т.е. все слова присутствуют, но смысл предложения отсутствует) заменяется переводом с более естественным звучанием.

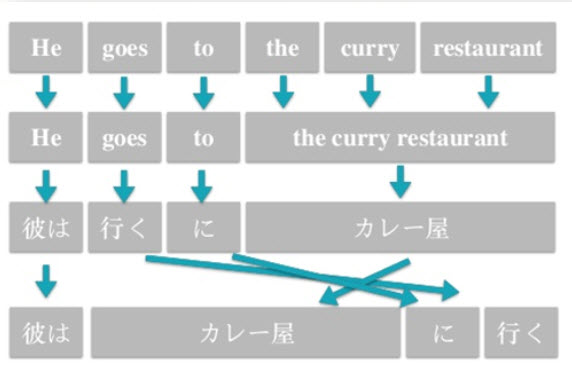
****

Рисунок 1 – Схема работы SMT

* SMT разбивает предложение на языке оригинала на фразы. (Вы можете увидеть фразу в виде отдельных серых блоков во второй строке диаграммы.)
* Затем SMT ищет каждую из этих фраз в таблице фраз / модели перевода и генерирует переводы на целевой язык. (Вы можете увидеть это в третьей строке диаграммы.)
* Затем опять же используется таблица фраз / модель перевода, чтобы изменить порядок этих фраз для улучшения качества перевода. (Вы можете увидеть это изменение порядка в четвертой строке диаграммы.)
  1. Выявление проблем SMT:
* Для создания корпусов требуются большие наборы данных (оригинального текста/корректно переведённого).
* Требуются большие объёмы памяти для хранения моделей перевода.
* При создании модели требуются постоянные корректировки от инженера для достижения более качественного перевода.
* Статический перевод не основывается ни на каких правилах языка из-за чего часто возникают ошибки в переводе.
* Нестабильность перевода из-за использования вероятностной модели, как основы для выбора подходящего перевода и из-за того, что корпуса уточняются эволюционно с помощью инженеров для улучшения перевода.
* Корпуса чаще всего должны быть ориентированы на какую-то конкретную предметную область для того, чтобы внутри модели слова использовались в каком-то конкретном, необходимом в данной предметной области значении и если пользователь выйдет за эту предметную область перевод будет с большой долей вероятности не корректным.
  1. Выявление преимуществ SMT:
* При наличии хорошо откорректированных корпусов с помощью SMT можно довольно точно переводить большие объёмы текста в заданной предметной области, требующие лишь небольших корректировок от пользователя..
* Просты в создании при наличии этих самых больших наборов данных (оригинального текста/корректно переведённого).

1. Анализ проблем и преимуществ RBMT:
   1. Описание работы RBMT:

Как и с прошлым видом MT начнём с определения. RBMT - система машинного перевода , основанная на лингвистической информации об исходных и целевых языках, в основном получаемой из (одноязычных, двуязычных или многоязычных) словарей и грамматик, охватывающих основные семантические, морфологические и синтаксические закономерности каждого языка соответственно. Имея входные предложения (на некотором исходном языке), система RBMT генерирует их для вывода предложений (на некотором целевом языке) на основе морфологического, синтаксического и семантического анализа как исходного, так и целевого языков, участвующих в конкретной задаче перевода. Если вкратце, то основной подход систем RBMT основан на связывании структуры заданного входного предложения со структурой требуемого выходного предложения, обязательно сохраняя их уникальное значение.

Далее проиллюстрируем общую структуру RBMT на примере:

Исходные данные:

* Исходный текст: A girl eats an apple.
* Язык оригинала: Английский.
* Язык на который мы хотим перевести исходный текст: Немецкий.

Необходимые инструменты:

* Словарь, который сопоставляет каждое английское слово с подходящим немецким словом.
* Правила, представляющие структуру регулярного английского предложения.
* Правила, представляющие структуру регулярного немецкого предложения.
* Правила, по которым можно связать эти две структуры вместе.

Этапы перевода:

* Получение базовой информации о части речи каждого исходного слова: a – артикль, girl – существительное, eats – глагол, an – артикль, apple – существительное.
* Получение синтаксической информации о глаголе «to eat»:  
  Present Simple, 3 лицо ед. число.
* Анализ исходной фразы:  
  an apple – объект для глагола eat.
* Перевод всех слов на немецкий:  
  a (артикль) => ein (артикль)  
  girl (существительное) => Mädchen (существительное)  
  eat (глагол) => essen (глагол)  
  an (артикль) => ein (артикль)  
  apple (существительное) => Apfel (существительное)
* Использование правил, сопоставляющих структуру и правила англ. и нем. языков, для корректного составления предложения на немецком:  
  A girl eats an apple. => Ein Mädchen isst einen Apfel.
  1. Выявление проблем RBMT:
* Недостаточное количество действительно хороших словарей. Создание новых словарей - дорогое удовольствие, и к тому же язык достаточно изменчив и словари могут устаревать.
* Трудно иметь дело с взаимодействием правил в больших системах, трудно переносить правила очень разнящихся по структуре и правилам языков, также существует проблема с неоднозначностью отношений слов в структуре предложения и неоднозначностью самих значений слов и не стоит забывать о идиоматических выражениях перевод которых не определяется значением входящих в них слов.
* Трудность адаптации к новым условиям. Хотя системы RBMT обычно предоставляют механизм для создания новых правил, расширения и адаптации лексикона, изменения обычно очень дороги, а результаты зачастую не окупаются.
  1. Выявление преимуществ RBMT:
* Не требуется никаких двуязычных текстов, как с SMT
* Нет качественного потолка. Каждую ошибку можно исправить с помощью целевого правила, даже если триггерный случай встречается крайне редко. Это контрастирует со статистическими системами, в которых редко встречающиеся формы по умолчанию стираются.
* Полный контроль. Поскольку все правила написаны от руки, вы можете легко отладить систему, основанную на правилах, чтобы точно увидеть, где данная ошибка попадает в систему и почему.
* Возможность повторного использования. Поскольку системы RBMT обычно строятся на основе сильного анализа исходного языка, намного проще написать новые правила для RBMT для перевода с языка с уже описанными правилами на новый.

1. Анализ проблем и преимуществ NMT:
   1. **Описание работы NMT:**

И наконец-то NMT – самая новая и многообещающая технология, которую ещё в 2016 году стали использовать Google на своём сервисе машинного перевода Google Translate, заменив нейронной сетью GNMT свои статические методы. Технически NMT мы можем называть все типы машинного перевода, использующие искусственные нейронные сети для реализации машинного перевода. Описать действие подобной нейронной сети можно так: каждое слово во входном предложении (например, английском) кодируется как число, которое нейронная сеть переводит в результирующую последовательность чисел, представляющую переведенное целевое предложение (например, китайское).  
Вот упрощённый пример работы NMT для перевода с английского на китайский :

* «Я собака» кодируется числами 251, 3245, 953, 2.
* Числа 251, 3245, 953, 2 вводятся в модель нейронного преобразования и приводят к выходным данным 2241, 9242, 98, 6342.
* 2241, 9242, 98, 6342 затем декодируется в китайский перевод   
  «我 是 只狗»

(каждое число на входе и выходе представляет слово в английском и китайском словаре и всегда кодируется и декодируется соответственно)

В приведенном выше примере возникает следующий вопрос: «Как работает модель перевода?» Простой ответ - через сложную математическую формулу (представленную в виде нейронной сети). Как описано ранее, эта формула принимает на входе строку чисел и выводит результирующую строку чисел. Параметры этой нейронной сети создаются и уточняются путем обучения сети миллионами пар предложений (например, перевод пар предложений на английский и китайский языки). Каждая пара предложений немного изменяет нейронную сеть, поскольку она проходит через каждую пару предложений с использованием алгоритма, называемого обратным распространением . В результате получается наиболее подходящая модель, наиболее точно переводящая любое из входных чисел в выходные числа из миллионов предоставленных пар предложений.

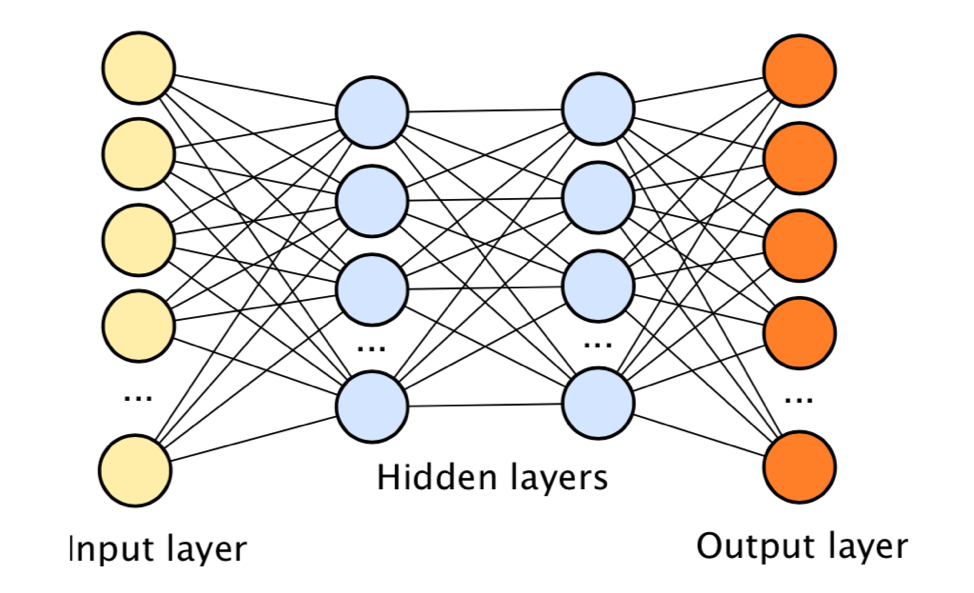


Рисунок - Визуализация нейронной сети

Проще говоря, нейронные сети допускают сложность. Нейронные сети могут иметь большое количество параметров с весами и смещениями между узлами, что дает им гибкость для соответствия очень сложным данным и обучения сложных моделей. Сложность модели позволяет обобщить ее на большие объемы примеров, на которых она обучается, например, для переваривания миллионов языковых пар.

* 1. Выявление проблем NMT:
* Поскольку нейронные сети обучаются за счёт больших объемов данных, их нелегко приручить и контролировать. Это становится особенно проблематичным, когда дело касается перевода имен (людей или компаний).
* Требуются большие вычислительные мощности и всё равно требуются переводчики-люди, чтобы контролировать обучение нейронной сети.
* Из-за того, что нейронные сети обучаются в основном переводить тексты не углублённые в какую-то определённую область знаний, то соответственно пытаясь перевести статью использующую термины из физики, юриспруденции или любой другой специализированной сферы знаний NMT будет работать плохо.
* Из-за этого же нейронные сети плохо справляются с редко встречающимися словами.
  1. Выявление преимуществ NMT:
* Нейронные сети не требуют от инженеров ни корпусов обучения, как того требует SMT, ни полностью прописанных правил перевода, как в RBMT, большую часть работы по составлению модели перевода и её усовершенствования делает сама нейронная сеть, требуя лишь контроля/ корректировок со стороны инженеров.

После проведения сравнительного анализа всех выше рассмотренных видов MT, был сделан вывод, что технологию SMT – лучше использовать для перевода узкой направленности, чтобы не возникало проблемы с неоднозначностью значений тех или иных слов/фраз, размер текста при этом нам не особо важен, RBMT – идеальный выход для перевода родственных или хорошо структурированных языков. С помощью актуальных и правильно сделанных словарей RBMT можно переводить корректно даже большие объёмы текста, а NMT вряд ли будет лучше для перевода узкоспециализированного текста чем SMT, не будет так контролируем, как словари правил RBMT, но в области коротких сообщений/статей, не углублённых в какую-то узкую специализацию, например, при переводе сообщений в online чатах NMT – будет лучшим выбором, при том NMT имеет самую низкую нагрузку на инженера из-за элемента самообучения.

ЗАКЛЮЧЕНИЕ

Итогом НИР является проанализированная проблема машинного перевода по анализу проблем и преимуществ различных типов машинного перевода. Решение содержит содержит описание каждого типа MT для понимания механизма их работы для выбора типа MT под свои задачи.

СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ