**УДК 811.111'322.4:34**

**Детскина Раиса Владимировна**
кандидат филологических наук, доцент
доцент кафедры информатики
и прикладной лингвистики
Минский государственный лингвистический
университет
г. Минск, Беларусь

**Raisa Detskina**
PhD in Philology
Associate Professor at Informatics
and Applied Linguistics Department
Minsk State Linguistic University
Minsk, Belarus
raisadetskina@gmail.com

# НЕЙРОСЕТЕВАЯ МОДЕЛЬ СИСТЕМЫ МАШИННОГО АНГЛО-РУССКОГО ПЕРЕВОДА ЮРИДИЧЕСКИХ ТЕКСТОВ

# NEURAL NETWORK MODEL OF THE ENGLISH-RUSSIAN MACHINE TRANSLATION SYSTEM OF LEGAL TEXTS

В данной статье излагаются наиболее важные аспекты нейросетевого моделирования систем машинного перевода текстов. Рассматриваются его суть и преимущества по сравнению с традиционным, статистическим методом решения задачи, приведены особенности реализации для системы машинного перевода юридических текстов с английского языка на русский. Даны примеры, иллюстрирующие качество получаемого системой выходного результата.

Ключевые слова: *входной текст*; *выходной текст*; *корпус текстов*; *машинный перевод*; *нейросетевая модель*; *машинное обучение*.

The article outlines the most important aspects of the neural network modeling of the machine translation systems. The essence and advantages of the presented model over the traditional statistical method of solving the tasks are considered in the article. The application features of the English-Russian machine translation system of legal texts are presented. The illustrative examples of the quality of the output result obtained by the system are provided.

Key words: *input text*; *output text*; *text corpus*; *machine translation*; *neural network model*; *machine learning*.

На начальных этапах развития искусственного интеллекта (ИИ) в соответствующих приложениях довольно эффективно решались задачи, интеллектуально сложные для людей, но относительно «прямолинейные» для компьютеров, которые могут быть описаны списком формальных математических правил. Позже оказалось, что гораздо сложнее обстоит дело с теми задачами, которые не могут быть представлены в формализованном виде, но, тем не менее, решаются нами относительно легко, т.е. задачами, которые человек решает интуитивно, как бы автоматически, такие, например, как распознавание слов или лиц на изображениях [1].

Именно этим обстоятельством и признанием того, что человеческий мозг обрабатывает информацию совершенно отличным от компьютера способом, было предопределено возникновение совершенно нового направления в решении многих задач ИИ и компьютерной лингвистики, основанного на машинном обучении и нейронных сетях. Мозг, образно говоря, сложный,

нелинейный и параллельный компьютер (система вычисления информации), который может организовывать структурные составляющие, известные как *нейроны*, для выполнения определенных вычислений и операций быстрее, чем любая из ныне существующих высокопроизводительных вычислительных машин.

Таким образом, мозг имеет сложную структуру и способность выстраивать свои собственные правила, что мы чаще всего называем *опытом*. Нейронная сеть – это своего рода машина, которая конструируется с целью моделирования того, как человеческий мозг осуществляет решение той или иной задачи или выполнение некоторой функции.

Понятие *искусственная нейронная сеть* впервые было введено в 40-х годах прошлого века. Исследователи в данной области до сих пор не пришли к единому мнению относительно определения нейронной сети, но наиболее цитируемым является определение, предложенное С. Хайкиным, которое отражает современный подход к пониманию нейронных сетей [2, с. 10]: «нейронная сеть – это процессор с массивно-параллельной архитектурой, состоящий из вычислительных элементов (нейронов), который имеет естественную склонность к запоминанию эмпирических знаний и представлению этих знаний в доступном виде». Нейронная сеть схожа с мозгом в двух отношениях:

- знания приобретаются нейронной сетью из окружающей среды посредством обучения;
- межнейронная сила связи, известная как синоптический вес, используется для хранения приобретенных знаний.

Процедура, применяемая для осуществления процесса обучения, называется *обучающим алгоритмом*, задача которого – упорядоченно модифицировать синоптические веса нейронной сети для достижения желаемой цели. Модификация синоптических весов обеспечивает традиционный метод для проектирования нейронной сети [1].

Современная эра нейронных сетей началась с новаторской работы У. Мак-Каллока и У. Питтса [3, с. 4]. В своей, уже классической, работе ученые описали логические вычисления нейронных сетей, которые объединили исследования нейрофизиологии и математической логики. Через 15 лет после выхода в свет фундаментальной работы У. Мак-Каллока и У. Питтса новый подход был предложен Ф. Розенблаттом, разработавшим модель персептрона для решения задачи классификации. В 1969 году М. Минский публикует формальное доказательство ограниченности перцептрона и показывает, что он неспособен решать некоторые задачи, связанные с инвариантностью представлений. На протяжении последующих 10 лет интерес к нейронным сетям резко упал, сопровождаясь публикациями таких известных ученых, как Т. Кохонен, Дж. Андерсон, Б. В. Хакимов, Пол Дж. Вербос, А. И. Галушкин, работы которых, тем не менее, не привлекли должного внимания. Этот период забвения продолжался вплоть до 1982 года, когда Дж. Хопфилд в своей работе доказал, что нейронная сеть с обратными

связями может представлять собой систему, минимизирующую энергию, что вызвало бурные волнения в научном мире. В 1986 году Д. И. Румельхартом, Дж. Е. Хинтоном и Р. Дж. Вильямсом и одновременно С. И. Барцевым и В. А. Охониным был существенно развит метод обратного распространения ошибки. С того момента и по сей день интерес к обучаемым нейронным сетям растет в геометрической прогрессии [1].

Таким образом, можно заключить, что такая отрасль, как глубокое обучение с использованием нейронных сетей, зародилась в начале 50-х годов XX века и претерпела большие изменения, то попадая, то уходя из поля зрения как самостоятельная область исследований. Ее появление было мотивировано фундаментальным убеждением, что человеческий мозг является самым мощным вычислительным механизмом, а прототипом искусственных нейронных сетей послужили биологические нейронные сети.

Способ, по которому нейроны в нейронной сети структурированы, тесно связан с алгоритмом обучения, используемым для тренировки нейронной сети. Таким образом, можно говорить об алгоритмах или правилах, используемых для проектирования нейронной сети, как о чем-то структурированном, то есть имеющих свою архитектуру. Существует несколько классификаций нейронных сетей по типу структуры (гомогенные, гетерогенные), типу сигнала (бинарные, аналоговые), по типу работы (синхронные, асинхронные), топологии (многослойные, однослойные) [3, с. 7]. Всего выделяют несколько основополагающих типов архитектуры нейронной сети [2, с. 5].

1. Однослойная нейронная сеть прямого распространения

В нейронной сети со слоями нейроны организованы в слои. В простейшей нейронной сети со слоями присутствует *входной слой входных узлов*, который проектируется на *выходной слой нейронов (вычисляемых узлов)*, причем исключительно в данном порядке. Другими словами, эта нейронная сеть строго ациклична.

2. Многослойная нейронная сеть прямого распространения

Второй тип искусственных нейронных сетей отличается наличием одного или более *скрытых слоев*, вычислительные узлы которых соответственно называются *скрытые нейроны* или *скрытые элементы*. Скрытые нейроны расположены между входным и выходным слоями. Посредством добавления одного или более скрытых слоев нейронная сеть способна извлекать статистику высокого порядка, что является практически ценным, когда размер входного слоя очень большой.

Нейросетевое моделирование не обошло своим вниманием и целый класс актуальных задач автоматической обработки текста, в том числе и задачи его машинного перевода (МП). Именно этому подходу были адресованы проблемы, существующие в рамках статистического машинного перевода: (1) игнорирование зависимостей, которые в тексте находятся на большом расстоянии друг от друга, и (2) комплексность, так как добавляется все больше и больше характеристик для улучшения работы систем статистического МП.

Нейросетевое моделирование в задаче машинного перевода – это одна большая нейронная сеть с миллионом искусственных нейронов, которые проектируются с целью моделирования всего процесса МП текста. Такой подход требует минимальных знаний предметной области, параллельного корпуса входных и выходных пар предложений, как и в статистическом машинном переводе, но с меньшим количеством вычислительных шагов, реализующих переводную модель. Одна из самых привлекательных особенностей нейросетевого моделирования в задаче МП – это то, что модель может быть обучена производить перевод сразу, без промежуточного обучения компонентов, в противовес статистическому машинному переводу [4].

В данной статье рассмотрено представление лингвистической информации в системе машинного перевода англоязычных юридических текстов на русский язык и, в частности, нейронное моделирование языка. Для создания соответствующей системы МП был задействован известный корпус параллельных текстов: United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. Данный корпус состоит из 2100 Резолюций Генеральной Ассамблеи ООН с переводом на 6 официальных языков Организации Объединенных Наций, из которых для проведения исследования была извлечена англо-русская языковая пара. В корпусе каждое английское предложение представлено с новой строки и имеет соответствующее ему предложение-перевод на русский язык. В рамках работы с данным корпусом были выбраны пары предложений, которые содержат не более 30 слов. В итоге для нашей задачи была сформирована выборка данных, равная 25 216 предложениям.

В практической реализации данной системы не использовались одноязычные данные кроме указанного параллельного корпуса, хотя может быть использован гораздо больший по объему англо-русский корпус.

Весь корпус тексов был разделен на тренировочную выборку (25 216 предложений), оценочную выборку (100 предложений) для подсчета качества перевода с текущими весами на промежуточном этапе и тестовую выборку (6 943 предложений) для итоговой оценки качества перевода.

В тренировочной выборке из множества всех входных и выходных предложений формируются словари уникальных слов: словарь уникальных слов выходного русского языка, содержащий 27 760 уникальных слов в 25 216 предложениях при общем количестве слов, равном 320 702, словарь уникальных слов входного английского языка, содержащий 19 061 уникальное слово в 25 216 предложениях при общем количестве слов, равном 365 493.

Тренировочная выборка, состоящая из англо-русских пар предложений, является основой для обучения данной нейросетевой модели машинного перевода. Для обработки и последующего представления лингвистической информации, в нашем случае слов, используется слой нейронной сети, который обеспечивает векторное представление входных и выходных значений. То есть каждое слово встроено в пространство заданной размерности и представлено в нем в виде вектора. Вложения – векторные представ-

ления каждого элемента – натренированы, как параметры функции внутри нейронной сети и в процессе обучения меняют координаты в пространстве, основываясь на встречаемости соответствующих слов в контексте. Для проведения исследования была выбрана размерность вектора, равная 128, так как опытным путем было выявлено, что его меньшая размерность дает худший результат, так как обладает меньшей способностью закодировать особенности такой лингвистической информации, как слово.

На этапе кодировки слов происходит построение двух языковых векторных моделей, которые отражают слова как точки, встроенные в некое многомерное пространство: входного английского текста и выходного русского текста.

Слова, которые не вошли в словари уникальных слов, заменяются специальным символом [UNK] и на этапе порождения перевода обрабатываются нейронной сетью отдельно, с использованием предварительно натренированных моделей представления слов GloVe: Global Vectors for Word Representation [1] и Word2Vec [5], которые находятся в открытом доступе и предоставляются их авторами для проведения исследований в области компьютерной лингвистики. Имплементация в систему дополнительных источников представления слов, естественно, может обогатить ее базу знаний и упростить процесс тренировки самой модели.

Стоит отметить тот факт, что благодаря вероятностному алгоритму, заложенному для вычисления векторных представлений, точки в их пространстве располагаются таким образом, что соответствующие им слова, которые являются семантически близкими, находятся в пространстве ближе друг к другу, и наоборот.

Другая важная особенность моделей состоит в том, что слова, которые являются переводными эквивалентами в английском и русском языках, имеют схожие значения векторов в английской и русской языковых моделях соответственно. Например, слово *unlawful* имеет вектор, равный 0,399, в модели входного языка, а слово *незаконный* имеет вектор, равный 0,348, в модели выходного языка.

И последняя характерная черта данных моделей, которая существенна для разработки целевой системы МП, – расстояния между векторами слов, которые вероятнее всего встретятся в контексте друг с другом, очень близки в двух векторных моделях. Как пример, расстояние между словами *unlawful* (0,399) и *act* (0,469) в модели входного языка равно расстоянию между словами *незаконное* (0,348) и *действие* (0,505) в модели выходного языка.

Таким образом, в нашей нейросетевой модели был использован слой нейронной сети, который определил векторное представление входных и выходных значений и сформировал две модели языка: для входного и выходного языков, причем, основываясь на контекстной встречаемости в корпусе. То есть каждое слово было встроено в пространство заданной размерности и представлено в нем в виде вектора.

Получившиеся нейросетевые модели отражают важные лингвистические особенности английского и русского языков: синонимичность, антонимичность их слов и контекстную зависимость между словами. Немаловажен тот факт, что данный способ представления удобен, так как генерирует вложения для большого корпуса текстов – больше 10 000 уникальных слов, работает с ненормализованными данными и представляет их в виде векторов заданной размерности.

Выбранная размерность – 128 – положительно отразилась на результатах. Так, при данном значении удалось сделать максимально репрезентативные языковые модели.

Были заданы следующие параметры, которые также повлияли на ход обучения модели: количество предложений, используемых для промежуточной оценки качества перевода (50), максимальное количество эпох – общее количество повторений, после которых обучение останавливается (40), размерность (128).

Для оценки качества МП текста по итогам обучения в рамках одной эпохи было использовано несколько общепринятых алгоритмов-метрик, которые приближены к человеческой оценке качества перевода: BLEU (bilingual evaluation understudy), CIDer, ROUGE_L, TER [6]. В нашем случае качество перевода есть степень соответствия выходного текста, полученного системой МП и полученного экспертом. На графике (рисунок) показан прогресс качества выходного результата системы МП англоязычных юридических текстов на русский язык в зависимости от степени обучения ее нейросетевой модели.

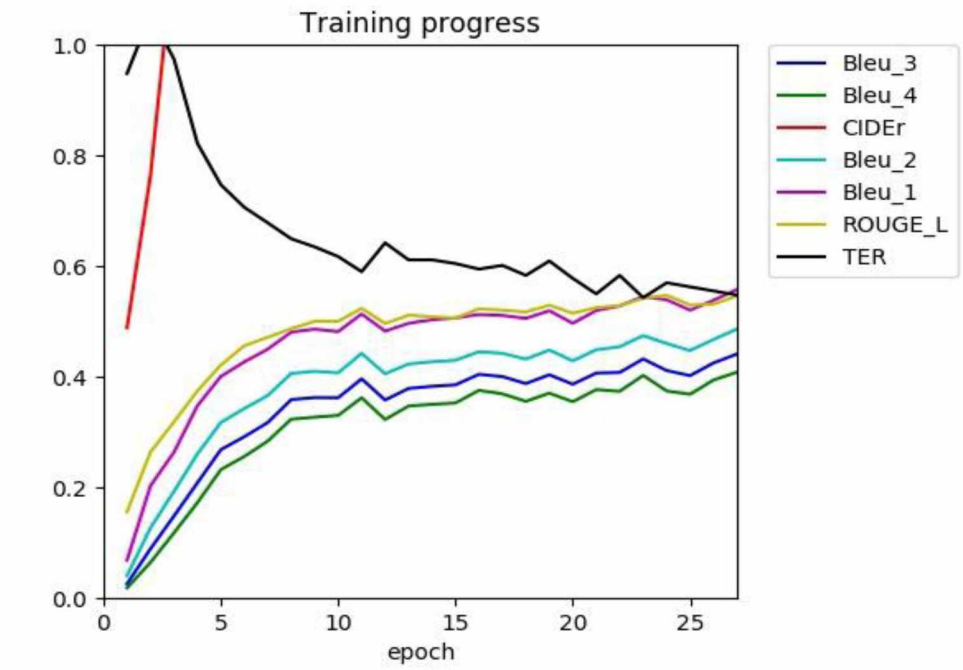


График зависимости качества МП англоязычных юридических текстов на русский язык: ось OY отражает качество перевода в диапазоне от 0 до 1, где 0 соответствует 0 %, а 1 – 100 %; ось OX показывает эпохи тренировки

Из графика можно сделать вывод, что нейронная сеть улучшила свою производительность в зависимости от степени обучения нейросетевой модели системы.

Выборочные результаты перевода предложений входного текста, полученные системой МП и, для сравнения, экспертом, представлены в таблице

Результаты перевода англоязычных юридических текстов на русский язык, полученные системой МП и экспертом

| № п/п | Предложения | Примеры |
|---|---|---|
| 1.1 | Входное предложение | *States Parties undertake to adopt immediate, effective and appropriate measures:*<br>*(a) To raise awareness throughout society, including at the family level, regarding persons with disabilities, and to foster respect for the rights and dignity of persons with disabilities.* |
| 1.2 | Машинный перевод | *Государства-участники обязуются принимать безотлагательные, эффективные и надлежащие меры и надлежащие меры к тому, чтобы:*<br>*а) повышать осведомленность общественности на всей общества, включая семьи и поощрять уважение прав человека и достоинства инвалидов.* |
| 1.3 | “Ручной” перевод | *Государства-участники обязуются принимать безотлагательные, эффективные и надлежащие меры к тому, чтобы:*<br>*а) повышать просвещенность всего общества, в том числе на уровне семьи, в вопросах инвалидности и укреплять уважение прав и достоинства инвалидов* |
| 2.1 | Входное предложение | *Recalling also the relevant Security Council resolutions.* |
| 2.2 | Машинный перевод | *Ссылаясь также на соответствующие резолюции Совета Безопасности.* |
| 2.3 | «Ручной» перевод | *Ссылаясь также на соответствующие резолюции Совета Безопасности.* |
| 3.1 | Входное предложение | *Encourages entities of the United Nations system to participate fully in the work of the Inter-Agency Meeting on Outer Space Activities* |
| 3.2 | Машинный перевод | *Рекомендует органам системы Организации Объединенных Наций в полной мере принимать участие в работе Межучрежденческого совещания по космической деятельности* |
| 3.3 | «Ручной» перевод | *Рекомендует органам системы Организации Объединенных Наций в полной мере принимать участие в работе Межучрежденческого совещания по космической деятельности* |

*Окончание таблицы*

| | | |
|---|---|---|
| 4.1 | Входное предложение | *Each State Party shall cooperate with the Committee and assist its members in the fulfilment of their mandate.* |
| 4.2 | Машинный перевод | *Каждое Государство-участник должны сотрудничать с Комитетом и оказывать помощь в выполнении ими своих мандата.* |
| 4.3 | «Ручной» перевод | *Каждое государство-участник сотрудничает с Комитетом и оказывает его членам содействие в выполнении ими своего мандата.* |

Заметим, что в целом качество перевода, полученное системой МП, в данном случае достаточно высокое. Существуют некоторые проблемы, связанные с переводом более сложных, длинных предложений, которые, исходя из сути нейросетевых моделей, преодолимы с использованием более объемных тренировочных корпусов текстов.

## ЛИТЕРАТУРА


1. *Pennington, J.* GloVe : Global Vectors for Word Representation / J. Pennington, R. Socher, Christopher D. Manning [Electronic resource]. – Stanford: Computer Science Department, Stanford Univ. – Mode of access : https://nlp.stanford.edu/pubs/glove.pdf. – Date of access : 14.05.2021.
2. *Хайкин, С.* Нейронные сети : полный курс / С. Хайкин. – 2-е изд. – М. : Вильямс, 2006. – 1104 с.
3. *Галушкин, А.* Нейронные сети. Основы теории / А. Галушкин. – М. : Горячая линия–Телеком, 2010. – 496 с.
4. *McCulloch, W. S.* A Logical Calculus of the Ideas Immanent in Nervous Activity / W. S. McCulloch, W. Pitts. – Bulletin of Mathematical Biophysics, 1943. – 133 p.
5. *Sutskever, I.* Sequence to Sequence Learning with Neural Network / I. Sutskever, O. Vinyals, Quoc V. Le // Cornell University Library [Electronic resource]. – Cornell Univ., 2014. – Mode of access : https://arxiv.org/abs/1409.3215. – Date of access : 10.04.2021.
6. *Mikolov, T.* Distributed Representations of Words and Phrases and their Compositionality / T. Mikolov [Electronic resource]. – Mode of access : https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-hrases-and-their-compositionality.pdf. – Date of access : 14.05.2021.


*Поступила в редакцию 03.06.2021*