УДК 81.25

МАШИННЫЙ ПЕРЕВОД. НЕЙРОПЕРЕВОД

© Р. Г. Мифтахова*, Е. А. Морозкина

Башкирский государственный университет
Россия, Республика Башкортостан, 450076 г. Уфа, ул. Заки Валиди, 32.

Тел.: +7 (917) 441 12 96.
**Email: miftahovar@yandex.ru*

Перевод с опорой на нейронные сети уверенно замещает статистический машинный перевод. В статье рассмотрен принцип работы нейронных сетей в переводе, методы «обучения» таких сетей, а также различия этих видов перевода. Статья раскрывает такие понятия, как «глубокое обучение», «вес», «кодирование», «декодирование», «вектор», «уровень», «встраивание слов» (word embeddings) и др. Обосновано предположение, что на данном этапе развития нейронных сетей в переводе более эффективным представляется их комбинирование со статистическими системами, т.е. использование гибридных систем машинного перевода, которые могли бы взаимодополнять друг друга и, таким образом, обеспечивать наиболее адекватный перевод фраз и предложений. В статье изучены и проанализированы возможности нейронного перевода, представлены модели обработки естественного языка в машинном переводе, описан принцип функциональности двух систем машинного перевода – статистического и на основе нейронных сетей, а также определены лингвистические особенности, преимущества и недостатки использования таких систем в процессе перевода.

Ключевые слова*: нейронный перевод, машинный перевод, двуязычный параллельный корпус, word2vec, skip-gram, CBOW.*

В современном мире новые технологии призваны не только облегчить жизнь человека, но и подвигнуть его к новым этапам технологического развития. Кроме того, «информационные технологии … приводят к новым способам употребления языковых форм» [3, с. 162]. Существенный скачок в развитии можно наблюдать в области машинного перевода, который совершенствуется день ото дня.

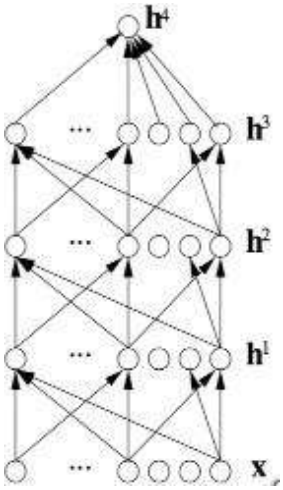
Актуальность сопоставления двух типов перевода, а именно машинного и нейроперевода, объясняется необходимостью дальнейшего усовершенствования процесса перевода. Новизна статьи обусловлена тем обстоятельством, что в ней изучаются технологии нейронных сетей, которые находятся на стадии внедрения в сферу перевода. Цель данной статьи заключается в сравнении двух систем машинного перевода, рассмотрении принципов работы нейропереводчика и моделей обработки естественного языка, которые в нем применяются.

Машинный перевод прошел долгий и сложный путь развития и превратился в незаменимого помощника, благодаря которому существенно сокращается как время, так и затраты на выполнение перевода. Заметным открытием в этой области стало внедрение нейронных связей в технологии машинного перевода. Это передовое изобретение, создатели которого черпали вдохновение из невероятно сложной работы нейронных сетей головного мозга, способно не просто обрабатывать и обобщать, но и анализировать информацию, а затем и учиться на основе полученных данных. Долгое время самым распространенным инструментом машинного перевода был статистический перевод, который по-прежнему остается востребованным, хотя и подвергается существенным модификациям. Однако последние несколько лет крупные компании, такие как Гугл, Майкрософт, а потом и Яндекс, начали применять технологии нейронных сетей в машинном переводе. Несмотря на все преимущества новой технологии перевода, на сегодняшний она не может полностью заменить уже существующие системы. Полагаем, что в настоящее время следует отдать предпочтение соотношению статистического метода с нейронными сетями для создания так называемых гибридных систем перевода. В такой системе две технологии взаимно компенсируют существующие в них недостатки.

Благодаря нейронным сетям существенно улучшилось качество машинного перевода. Нейросетевой машинный перевод не просто ищет и сопоставляет слова и выражения двуязычных корпусов, с его помощью становится возможным глубже проникнуть в связи, существующие между словами, и путем сложного анализа каждого переводимого образца изучить их взаимоотношения для выяснения контекста. Например, предложение “Can I have some tea?” статистическим методом будет переведено как «Я могу пить чай?», а нейросетевым – «Можно чаю?» Нейронная сеть анализирует «окружение» каждого слова, даже если слова находятся в разных частях предложения (Skip-grams), что позволяет правильно определить контекст, который, в свою очередь, обеспечивает связь слов в предложении при переводе. Нейронной сети необходимо время, чтобы самообучиться. На начальном этапе могут выдаваться несогласованные фразы: «Петр пошла» или «заманчивый идея». Такие проблемы решаются за счет статистического метода, в частности, использованием модели языка, которая содержит свод знаний о языке, выраженный в веро-

ятностном значении. Нет необходимости программировать нейронную сеть, такая сеть учится на основе примеров или же методом проб и ошибок. Одним из наглядных результатов подобного обучения может служить способность нейросетей обобщать и соотносить данные. После успешного обучения нейронная сеть может найти логичное решение для проблем одного типа, с которыми она до этого не сталкивалась [7, с. 3]. Например, «какао» и «кофе» часто появляются в сходных контекстах, а значит, оба эти слова должны быть возможны в контексте нового слова «пить», хотя в обучающих данных с этим словом, допустим, встречалось только слово «кофе». Таким образом, в результате использования нейроперевода появляется фраза «пить какао».

Структуру нейроперевода можно представить следующим образом:



x – необработанный сигнал (звук, слово, символ и т.д.) – входные данные;
h_1 - внешний слой;
h_2 и **h_3** – скрытые (внутренние) слои;
h_4 – выходные данные.

Скрытые слои содержат сложные математические вычисления, позволяющие определить контекст, а перевод в такой системе означает кодирование и декодирование. Представим, что необходимо перевести фразу “His flight is delayed” на русский язык. Репрезентация предложения формируется из векторных вложений отдельных слов. Кодировщик на основе имеющихся знаний, а точнее, корпусных данных, оценивает, может ли “His” быть в начале предложения, затем рассчитывает вероятность того, что после “His” может следовать “flight”, и т.д. Затем происходит кодировка справа налево, как если бы предложение выглядело как “delayed is flight His”. Следующий этап представляет собой непосредственно перевод. Декодер определяет наиболее вероятностный перевод для слова “His”, стоящего в начале предложения («Его»), затем – для “His flight” («Его рейс») и т.д. В итоге, для исходных фраз “His flight” и “His flight is” перевод окажется одинаковым, и на выходе сформируется предложение «Его рейс задерживается». Технологии глубокого обучения, применяемые в сфере языка, напрямую связаны со значением слова (word meaning), которое принимается за вектор чисел [12]. «Для создания адекватного перевода следует использовать комплексные модели перевода… включающие в себя элементы семантической, ситуативной и коммуникативной моделей [4, с. 132]. Существуют два уровня сходства слов: внешний (по форме) и семантический (по значению). Например, слова «бык», «корова» и «теленок» имеют между собой мало общего, однако на семантическом уровне очевидно, что они используются для выражения рода и возраста одного и того же вида животных. Другой пример относится к внешне (по форме) сходным словам: например, слова «моросить» и «морозить». Так, для удобства представления семантического сходства слов была предложена так называемая модель *векторного представления слова (word embedding)*, которая помогает сопоставить слово с определенным, присущим только ему вектором во всем пространстве смыслов. Вектор показывает коннотацию слова в числовом виде, а также то, что это слово помещено в многомерное векторное пространство. Используя методы для изучения векторов слов, применяемые в глубоком обучении (Deep learning), и помещая эти слова в многомерные векторные пространства, можно спровоцировать их работу в качестве семантически связных слов. Лексические единицы с похожими значениями также группируются вместе в векторном пространстве. При изучении вектора слова требуются многомерные векторные пространства (300-мерные, 1000-мерные и т.д.). Каждому вектору соответствует своя ось. Однако значение слова не всегда лежит вдоль определенной оси, на самом деле оно может лежать под любым другим углом в векторном пространстве. Например, самыми близкими значениями к слову “frog” будут: frogs, toad, litoria, leptodactylidae, rana, lizard, eleutheutherodactylus. Получается, что слова “frogs” и “toad” являются самыми близкими по значению слову “frog”.

Долгие годы исследователи по-разному подходили к вопросу определения семантического значения слов. В основном для этих целей использовали таксономические ресурсы. Самым известным среди них является электронный тезаурус и семантическая сеть WordNet. Несмотря на то, что WordNet, который долгое время широко применялся на практике, является большой базой данных смыслов различных слов, существует много причин, по которым из него трудно извлечь необходимые данные. Одна из причин заключается в том, что на этом уровне таксономических отношений теряется огромное количество нюансов. Например, среди синонимов слова “good” есть «adept», «expert», «good», «practiced» и т.д. Хотя в действительности все эти слова имеют разные значения.

Вектор слова в нейросети вычисляется иначе, чем с использованием WordNet. Все многообразие слов представлено в большом пространстве смыслов, которое, как упоминалось выше, может быть и 1000-мерным, и 5000-мерным и т.д. Вектор является измерением слова в пространстве смысла. Например, условно, пространство смысла состоит из 7-ми тем: политика, экономика, юриспруденция, молодежь, общество, здравоохранение, образование. И есть слово “expect”. Высчитывается вероятность, с которой данное слово встречается в корпусах каждой тематики (количество примеров данного слова в корпусе делится на общее число слов). Допустим, получаются следующие значения: в политической теме слово “expect” встречается с вероятностью P(0,286), в экономике – P(0,792), в юриспруденции – P(0,101), молодежь – P(0,567), общество – P(0,985), здравоохранение – P(0,673), образование – P(0,898).

$$\text{Expect} = \begin{pmatrix} 0.286 & =0 \\ 0.792 & =0 \\ 0.101 & =0 \\ 0.567 & =0 \\ 0.985 & =1 \\ 0,.673 & =0 \\ 0.898 & =0 \end{pmatrix}$$

Для дальнейшей обработки и передачи информации эти данные переводятся в более удобную двоичную форму – one-hot encoding, так как подавляющее большинство статистических систем и систем, основанных на правилах обработки естественного языка (Natural Language Processing – NLP) работают со словами как с «атомными» символами. В векторном пространстве атомный символ – это вектор, у которого все позиции равны 0 везде, кроме одной, равной 1. В языке существует большое количество слов, эквивалентных данным символам, поэтому 1 помещается в ту позицию вектора, которая представляет конкретный символ. Длина кода зависит от количества кодируемых объектов, в данном случае их 7, следовательно, получаем условный вектор данного слова: expect = [0000100]. А если пространство многомерное и каждое слово в этом пространстве имеет четкое измерение, то, естественно, таких значений будет довольно много.

Проблема заключается в том, что такая двоичная модель не дает представления о внутренних отношениях между словами, сходстве слов, тогда как необходимо знать, в каких случаях значения, слова и фразы схожи. Например, при машинном переводе “Seattle motel”, система в корпусе должна искать также фразу “Seattle hotel”. Вектором определяется не только отдельно взятое слово, но и фраза или предложение. В этом состоит их основное, но далеко не главное отличие от статистического машинного перевода. Векторное окружение слова позволяет понять основную суть всего документа, находить семантические и синтаксические сходства, а также связи с другими словами. Word embedding представляет слова в цифрах и сопоставляет их. Одним из популярных методов анализа и изучения семантики естественных языков с использованием нейронной сети является Word2Vec, основанный на том, что слова, которые встречаются в одном окружении, имеют общее семантическое поле. Word2Vec рассчитывается двумя способами: Skip-grams и Common Bag of Words (CBOW).

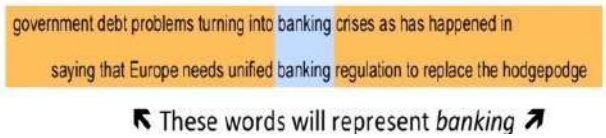
Рассмотрим схожие по значению предложения: “Have a good day” и “Have a great day”. Если создать для них свой словарь (V – vocabulary), который будет векторным пространством, то получится следующее: V = {Have, a, good, great, day}. Для каждого из этих слов в пространстве V, создается вектор с двоичным кодом. Длина вектора будет равна размеру V, т.е. 5. Получается следующее:

Have = [10000]; a = [01000]; good = [00100];
great = [00010]; day = [00001].

Если представить, что эти коды существуют в 5-мерном пространстве, где каждое слово занимает одно из измерений (без проекции на другие измерения), то оно не будет иметь ничего общего с другими словами. Это значит, что слова «good» и «great» будут так же различны, как слова «day» и «have», что не вполне соответствует действительности, ведь основная цель состоит в том, чтобы слова со схожим контекстом занимали близкие пространственные позиции.

Модель Skip Gram основана на дистрибутивном сходстве, когда представление слова кодирует его значение таким образом, что сходство между словами легко прослеживается. Понятие дистрибутивного сходства позволяет получить оценку для представления значения слова при обращении к контексту, в котором это слово появляется, т.е. задав конкретное слово в середине предложения (входное слово), система анализирует слова, стоящие рядом с ним и выбирает любое из них. Слова, окружающие главное слово, называются параметром окна. Типичный размер окна может доходить до 5 элементов, то есть 5 слов «до» и 5 слов «после» (всего 10). Далее сеть рассчитает вероятность каждой n-граммы и так называемой скип-граммы, за исключением стоп-слов. Британский лингвист Дж.Р. Ферс выразил это так: «You shall know a word by the company it keeps», т.е. значение слова определяется его контекстом. Возьмем, к примеру, слово “banking”. Если необходимо выяснить значение этого слова, нужно найти тысячи примеров, где это слово употребляется в текстах, чтобы посмотреть на среду, на окружение, в котором это слово появляется. Например, “debt problems”, “crisis”, “Europe” и т.д. Все выявленные слова подсчитываются с учетом контекста, чтобы получить представление о значении искомого слова. Таким образом, если предсказать, в каком контексте появится искомое слово, то его значение будет не так

сложно вычислить. Для каждого искомого слова, будет вычисляться его вектор, затем данный вектор будет использоваться для определения вероятности использования других слов, которые появляются в контексте искомого слова [12].



Переход к использованию нейронных сетей начался в Yandex в сентябре 2017 г. Нейронная сеть работает с большими параллельными корпусами, анализирует их с целью обнаружения закономерностей, на которых потом учится. Нейросеть не только переводит предложения максимально близко к естественному языку, но и производит определенную проверку, проводит каждое слово через модель языка во избежание несогласования и устранения любых грамматических или орфографических ошибок [18].

Несмотря на все очевидные достоинства нейронной сети в переводе, существуют и некоторые недостатки. При переводе незнакомого фрагмента текста нейросеть либо выбросит его, либо начнет «придумывать» что-то свое и будет ждать, пока это слово окончательно войдет в широкое употребление. Особенно это касается недостаточно распространенных имен или названий, редких слов и выражений, сленга или намеренного искажения слов, которое, например, часто встречаются в заголовках газет. Самый известный пример, получивший широкое распространение в сети, это перевод простого предложения: “Good morning kids!” как «Хорошие морники!», где система не смогла корректно распознать синтаксис, посчитав слово «morning» определением к слову «kids», и «выдумала» новое слово. Другая проблема состоит в том, что для обучения нейронной сети требуются корпуса больших объемов, чем для статистической системы (до 500 млн, чтобы обеспечить надлежащий перевод), а следовательно, требуется чрезвычайно большая вычислительная мощность [18].

Несмотря на все преимущества, нейронному переводчику необходимо время для самообучения. Нейронный переводчик обрабатывает целые предложения, а статистический – делит его на n-граммы, при этом нейронный переводчик не может рассчитать модель языка, тогда как статистический может. Таким образом, полагаем, что на данной стадии развития нейронных сетей в переводе наиболее продуктивным является использование гибридных систем, с помощью которых можно компенсировать недостатки двух разных подходов к переводу [19–20].

ЛИТЕРАТУРА

1. Алимов В. В., Артемьева Ю. В. Специальный перевод: практический курс перевода: изд. 2. М.: ЛЕНАНД, 2017. 208 с.
2. Злобин В. К., Ручкин В. Н. Нейросети и нейрокомпьютеры: учеб. пособие. СПб.: БВХ – Петербург, 2011. 256 с.
3. Морозкина Е. А., Мифтахова Р. Г. Влияние информационных технологий на развитие лингвистических норм. Башкирский гос. ун-т. Вестник, 2012. №1. С. 162–164.
4. Морозкина Е. А., Мифтахова Р. Г. Основы моделирования классического машинного перевода и статистического машинного перевода. Мат-лы конф. «Актуальные проблемы контрастивной лингвистики, типологии языков и лингвокультурологии в полиэтническом пространстве», 2011. С. 130–135.
5. Морозкина Е. А., Мифтахова Р. Г Основы моделирования классического машинного перевода и статистического машинного перевода. Актуальные проблемы контрастивной лингвистики, типологии языков и лингвокультурологии в полиэтническом пространстве. Сб. науч. ст. в 3-х ч. Мин. обр. и науки РФ, Башкирский гос. ун-т. 2011. С. 130–135.
6. Стейнбек Дж. Гроздья гнева / пер. Н. Волжина. М., Азбука. 2014. 607 с.
7. Kriesel D. A Brief Introduction to Neural Networks. 2007. 286 p.
8. Steinbeck John The Grapes of Wrath. 2014. 678 p.
9. URL: https://help.smartcat.ai/hc/ru/articles/360017711291-16-03-2017-
10. URL: https://www.andovar.com/machine-translation/
11. URL: https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing
12. URL: http://nautil.us/issue/21/information/the-man-who-tried-to-redeem-the-world-with-logic
13. URL: https://ain.ua/2017/03/03/kak-rabotayut-nejroseti/
14. URL: https://ai-science.ru/dzhon-dzhozef-xopfild/
15. URL: http://users.ics.aalto.fi/teuvo/
16. URL: https://hi-news.ru/research-development/iskusstvennyj-intellekt-i-dzheffri-xinton-otec-glubokogo-obucheniya.html
17. URL: https://habr.com/post/330654/
18. URL: https://yandex.ru/blog/company/kak-pobedit-mornikov-yandeks-zapustil-gibridnuyu-sistemu-perevoda
19. URL: https://www.andovar.com/machine-translation/ Stanford University School of Engineering, Lecture 2: Word Vector Representations: word2vec
20. URL: https://www.youtube.com/watch?v=ERibwqs9p38&index=2&list=PL3FW7Lu3i5Jsnh1rnUwq_TcylNr7EkRe6
21. URL: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa
22. URL: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
23. URL: https://www.microsoft.com/en-us/translator/business/machine-translation/#nnt
24. URL: https://habr.com/ru/post/414343/
25. URL: https://yandex.ru/company/press_releases/2011/0316/
26. URL: https://www.eduherald.ru/ru/article/view?id=18262
27. URL: https://yandex.ru/company/technologies/translation/
28. URL: https://nplus1.ru/material/2016/11/04/recurrent-networks
29. URL: http://www.promt.ru/translation_software/home/master/#PHP0003#
30. URL: https://www.bbc.com/news/world-asia-47018747
31. URL: https://www.bbc.com/russian/news-47012916
32. URL: https://lim-english.com/posts/tehnicheskii-angliiskii/

Поступила в редакцию 27.05.2019 г.

MACHINE TRANSLATION. NEURAL TRANSLATION

© R. G. Miftakhova*, E. A. Morozkina

Bashkir State University
32 Zaki Validi Street, 450076 Ufa, Republic of Bashkortostan, Russia.

Phone: +7 (917) 441 12 96.
**Email: miftahovar@yandex.ru*

Neural networks are slowly but surely replacing statistical machine translators. The authors of the article describe the principle of neural networks in translation, methods of their training, as well as the main differences with statistical translation systems. The article reveals such concepts as “deep learning”, “weight”, “coding”, “decoding”, “vector”, “level”, “word embedding”, and others. The assumption is proved, that at the present stage of development of neural networks in translation, its combination with statistical systems is more effective, i.e. the use of hybrid machine translation systems that could complement each other and thus provide the most adequate translation of phrases and sentences. The authors study and analyze the possibilities of neural translation, show the models of natural language processing in machine translation, describe the principle of the functionality of two systems – statistical and neural networks, as well as linguistic features, advantages, and disadvantages of using such systems in the translation process.

Keywords: Statistical Machine Translation, parallel corpus, neural translation, word2vec, skip-gram, CBOW.

Published in Russian. Do not hesitate to contact us at bulletin_bsu@mail.ru if you need translation of the article.

REFERENCES

1. Alimov V. V., Artem'eva Yu. V. Spetsial'nyi perevod: prakticheskii kurs perevoda: izd. 2 [Special translation: practical translation course: 2nd Ed.]. Moscow: LENAND, 2017.
2. Zlobin V. K., Ruchkin V. N. Neiroseti i neiro-komp'yutery: ucheb. posobie [Neural networks and neurocomputers: textbook]. Saint Petersburg: BVKh – Peterburg, 2011.
3. Morozkina E. A., Miftakhova R. G. Vliyanie informatsionnykh tekhnologii na razvitie lingvisticheskikh norm. Bashkirskii gos. un-t. Vestnik, 2012. No. 1. Pp. 162–164.
4. Morozkina E. A., Miftakhova R. G. Osnovy modelirovaniya klassicheskogo mashinnogo perevoda i statisticheskogo mashinnogo perevoda. Mat-ly konf. «Aktual'nye problemy kontrastivnoi lingvistiki, tipologii yazykov i lingvokul'turologii v polietnicheskom prostranstve», 2011. Pp. 130–135.
5. Morozkina E. A., Miftakhova R. G Osnovy modelirovaniya klassicheskogo mashinnogo perevoda i statisticheskogo mashinnogo perevoda. Aktual'nye problemy kontrastivnoi lingvistiki, tipologii yazykov i lingvokul'turologii v polietnicheskom prostranstve. Sb. nauch. st. v 3-kh ch. Min. obr. i nauki RF, Bashkirskii gos. un-t. 2011. Pp. 130–135.
6. Steinbek J. Grozd'ya gneva [The grapes of wrath] / per. N. Volzhina. M., Azbuka. 2014.
7. Kriesel D. A Brief Introduction to Neural Networks. 2007.
8. Steinbeck J. The Grapes of Wrath. 2014.
9. URL: https://help.smartcat.ai/hc/ru/articles/360017711291-16-03-2017-
10. URL: https://www.andovar.com/machine-translation/
11. URL: https://www.frontiersin.org/research-topics/4817/artificial-neural-networks-as-models-of-neural-information-processing
12. URL: http://nautil.us/issue/21/information/the-man-who-tried-to-redeem-the-world-with-logic
13. URL: https://ain.ua/2017/03/03/kak-rabotayut-nejroseti/
14. URL: https://ai-science.ru/dzhon-dzhozef-xopfild/
15. URL: http://users.ics.aalto.fi/teuvo/
16. URL: https://hi-news.ru/research-development/iskusstvennyj-intellekt-i-dzheffri-xinton-otec-glubokogo-obucheniya.html
17. URL: https://habr.com/post/330654/
18. URL: https://yandex.ru/blog/company/kak-pobedit-mornikov-yandeks-zapustil-gibridnuyu-sistemu-perevoda
19. URL: https://www.andovar.com/machine-translation/ Stanford University School of Engineering, Lecture 2: Word Vector Representations: word2vec
20. URL: https://www.youtube.com/watch?v=ERibwqs9p38&in-dex=2&list=PL3FW7Lu3i5Jsnh1rnUwq_TcylNr7EkRe6
21. URL: https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa
22. URL: http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/
23. URL: https://www.microsoft.com/en-us/translator/business/machine-translation/#nnt
24. URL: https://habr.com/ru/post/414343/

25. URL: https://yandex.ru/company/press_releases/2011/0316/
26. URL: https://www.eduherald.ru/ru/article/view?id=18262
27. URL: https://yandex.ru/company/technologies/translation/
28. URL: https://nplus1.ru/material/2016/11/04/recurrent-networks
29. URL: http://www.promt.ru/translation_software/home/master/#PHP0003#
30. URL: https://www.bbc.com/news/world-asia-47018747
31. URL: https://www.bbc.com/russian/news-47012916
32. URL: https://lim-english.com/posts/tehnicheskii-angliiskii/

Received 27.05.2019.