

Разработка метода автоматического перевода пиктограммного сообщения в русскоязычный текст на основе машинного обучения

Д.С. Матюшечкин, А.Р. Донская

Волгоградский государственный технический университет

Аннотация: В статье описан метод для перевода пиктограммных сообщений в связный текст на русском языке на основе машинного обучения. В тексте статьи приведены понятие систем альтернативной коммуникации, описание процесса подготовки данных для обучения нейронной сети, описание используемой архитектуры и приведены результаты обучения.

Ключевые слова: системы альтернативной коммуникации, машинный перевод, машинное обучение, нейронная сеть, архитектура трансформера, Python, программное обеспечение.

Введение

На данный момент стремительно растет число людей, у которых наблюдаются нарушения речи, которые, как правило, вызваны психофизическими или ментальными расстройствами. Речевые нарушения могут вызывать осложнения как коммуникативной, так и познавательной функции речи, что подтверждают данные Всемирной Организации Здравоохранения как по России, так и по всему миру [1]. Таким людям присущи проблемы с использованием вербальной коммуникации, что впоследствии приводит к значительному снижению социальной активности.

Для того, чтобы улучшить социальную адаптацию людей, имеющих как нарушения речи, так и ментальные расстройства, обычно применяются средства альтернативной коммуникации, позволяющие частично, а в некоторых случаях и полностью заменить устную речь, что способствует социальной активности и помогает людям в повседневной жизни, будь то обучение, работа или общение [2].

Изучив ряд исследований, мы сделали заключение, что наиболее удобным и часто используемым средством для альтернативной коммуникации у людей с нарушениями речи является общение с помощью

цветных или черно-белых пиктограмм (изображения, которые помогают заменить слова) [3].

На данный момент разработано множество коллекций пиктограмм, помогающих заменить обычную речь, с помощью которых реализуется возможность задавать вопросы, выражать просьбы и так далее. Одними из наиболее известных коллекций пиктограмм являются PECS (Picture exchange communication system), Sclera, Beta. Однако, такие карточки не всегда могут быть верно интерпретированы или использованы без специальной подготовки, из чего следует необходимость в разработке методов для перевода текста в набор пиктограмм и обратно, для облегчения коммуникации между людьми с нарушениями речи и людьми, использующими вербальную коммуникацию.

Анализ существующих электронных средств для облегчения процесса коммуникации людей с речевыми нарушениями

На сегодняшний день существуют различные электронные средства для облегчения процесса коммуникации людей с речевыми нарушениями. Однако не существует средства, которое дает возможность переводить пиктограммные сообщения в связный текст на русском языке. В процессе исследования был проведен анализ существующих российских и зарубежных технологий в области автоматического перевода систем альтернативной коммуникации на естественный язык.

Для перевода на русский язык существуют разработки, выполненные на территории РФ, которые позволяют переводить последовательность пиктограмм в последовательность инфинитивов на русском языке без согласования частей речи, словоформ с учетом предлогов и учета других правил русского языка. Такие системы могут подойти для базовой коммуникации и обмена информацией, но подобные переводчики не

подходят для обучения пиктограммным языкам, например, с помощью озвучивания полученных фраз. Более того, озвучивание неверно согласованных предложений может усугубить процесс развития речи и сказаться негативно на пользователе [4].

Также были рассмотрены зарубежные проекты, связанные с переводом на английский, голландский и испанский языки. Для этих языков существуют переводчики, которые производят согласование слов в предложении, но не позволяют расширять банк пиктограмм и словарь новыми концептами. В основе изученных проектов лежат эвристические правила языков, грамматика которых значительно проще, чем грамматика русского языка. Следовательно, переводчики с иностранными языками не применимы для русскоязычных пользователей, а методы, которые используются в основе переводчиков, не применимы для перевода на русский язык.

В процессе данного исследования была разработана модель нейронной сети, которая позволит достаточно точно согласовывать слова в предложении с учетом возможных нарушений порядка слов и отсутствия некоторых предлогов, союзов и местоимений, чтобы результаты перевода можно было прослушивать и использовать как для коммуникации, так и для обучения, без дополнительной обработки результатов перевода пиктограммного сообщения в текст.

Описание метода автоматического перевода пиктограммного сообщения в русскоязычный текст на основе машинного обучения

Концепты языка систем альтернативной коммуникации представляются графическими изображениями (пиктограммами). Как правило, одна пиктограмма представляет собой одно слово, составляя последовательности которых, можно строить сообщения.

В процессе анализа задачи перевода пиктограммных сообщений в связный текст на русском языке было выявлено, что для каждой

используемой пиктограммы известна ее текстовая интерпретация на русском языке в инфинитивной форме. Вследствие этого, любое пиктограммное сообщение можно представить в виде текстовой последовательности инфинитивов (лемм). Тогда задача перевода пиктограммного сообщения заключается в согласовании последовательности лемм в связный текст на русском языке.

Задача согласования последовательности лемм в связное русскоязычное предложение схожа с задачей машинного перевода текста с одного языка на другой. Пример перевода пиктограммного сообщения, составленного из четырех пиктограмм в нотации «Beta» со значениями «я», «хотеть», «новый» и «игрушка» в связный текст на русском языке, приведен на рисунке 1.

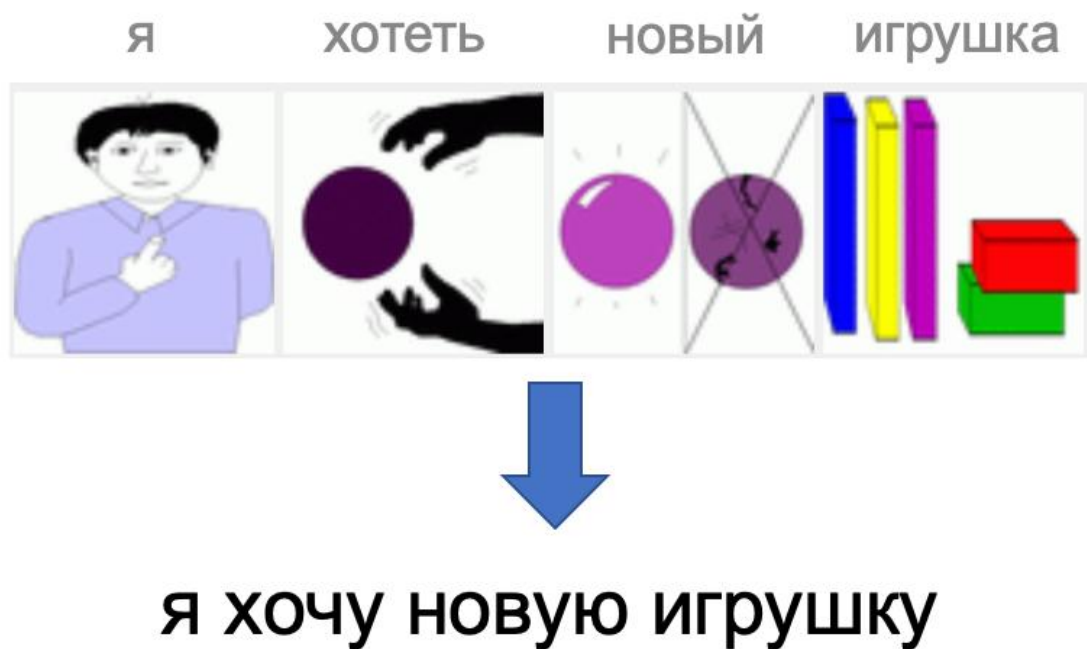


Рис. 1. – Пример перевода пиктограммного сообщения «Я хочу новую игрушку»

На данном примере видно, что сообщение на исходном языке – это последовательность лемм «я хотеть новый игрушка», а сообщение на целевом языке – это согласованное предложение «Я хочу новую игрушку».

Для машинного перевода с одного языка на другой применяются методы машинного обучения на параллельных текстовых корпусах, которые должны представлять собой одинаковые тексты на разных языках [5].

Тогда для задачи перевода пиктограммного сообщения в русскоязычный текст нужны такие параллельные текстовые корпуса, где исходный корпус содержит последовательности лемм, а целевой текстовый корпус содержит связные предложения, составленные из лемм исходного текстового корпуса.

Подготовка таких текстовых корпусов была произведена в несколько этапов:

1) Сбор целевого корпуса связных предложений на русском языке. Сбор производился путем автоматического разбора русскоязычных веб-страниц ресурса «Википедия», в результате чего был собран 1 миллион предложений.

2) Лемматизация всех слов в предложениях целевого корпуса для получения исходного текстового корпуса, состоящего из последовательностей лемм. Для лемматизации каждого слова в согласованных предложениях был использован инструмент “Яндекс-стеммер” для языка программирования “Python” [6].

3) Разбиение подготовленных корпусов на две части, чтобы по 800 тысяч предложений каждого корпуса были использованы для обучения нейронной сети и по 200 тысяч пар для оценки ее точности [7].

Однако в силу того, что последовательность лемм на входе переводчика образуется путем составления фраз из пиктограмм, то в таких последовательностях часто опускаются предлоги или путается порядок изображений. Например, для составления предложения «Я хожу в школу.» используются только пиктограммы со значениями «я», «ходить» и «школа», а правильный предлог при переводе должен быть добавлен нейронной сетью.

Поэтому дополнительно в исходный текстовый корпус с последовательностями лемм был добавлен «статистический шум». В исходном текстовом корпусе было произведено удаление одной десятой части предлогов, случайно переставлены местами леммы и продублированы некоторые последовательности лемм.

После подготовки текстов было произведено обучение нейронной сети для создания переводчика.

Описание архитектуры нейронной сети

В качестве нейросетевой модели перевода последовательностей инфинитивных форм в согласованные предложения была использована архитектура трансформера, которую представили в статье «Attention is all you need» авторы из компании «Google» в 2017 году. Модель нейронной сети трансформера представлена модулями кодера и декодера.

Кодер получает на вход векторное представление входной последовательности (Embeddings) и векторное представление номеров слов (positional embedding). Декодер получает на вход часть этой последовательности и выход кодера. Кодер и декодер состоят из слоев. Слои кодера последовательно передают результат следующему слою в качестве его входа. Слои декодера последовательно передают результат следующему слою вместе с результатом кодировщика в качестве его входа.

Каждый кодировщик состоит из механизма самовнимания (вход из предыдущего слоя) и нейронной сети с прямой связью (вход из механизма самовнимания). Каждый декодер состоит из механизма самовнимания (вход из предыдущего слоя), механизма внимания к результатам кодирования (вход из механизма самовнимания и кодировщика) и нейронной сети с прямой связью (вход из механизма внимания).

Каждый механизм внимания параметризован матрицами весов запросов W_Q , весов ключей W_k и весов значений W_V . Для вычисления внимания входного вектора X к вектору Y , вычисляются векторы $Q = W_Q X$, $K = W_K X$, $V = W_V Y$ [8].

Эти векторы используются для вычисления результата внимания по формуле:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

где d_k - характеристика размера ключей.

Схематичное изображение архитектуры трансформера представлено на рисунке 2.

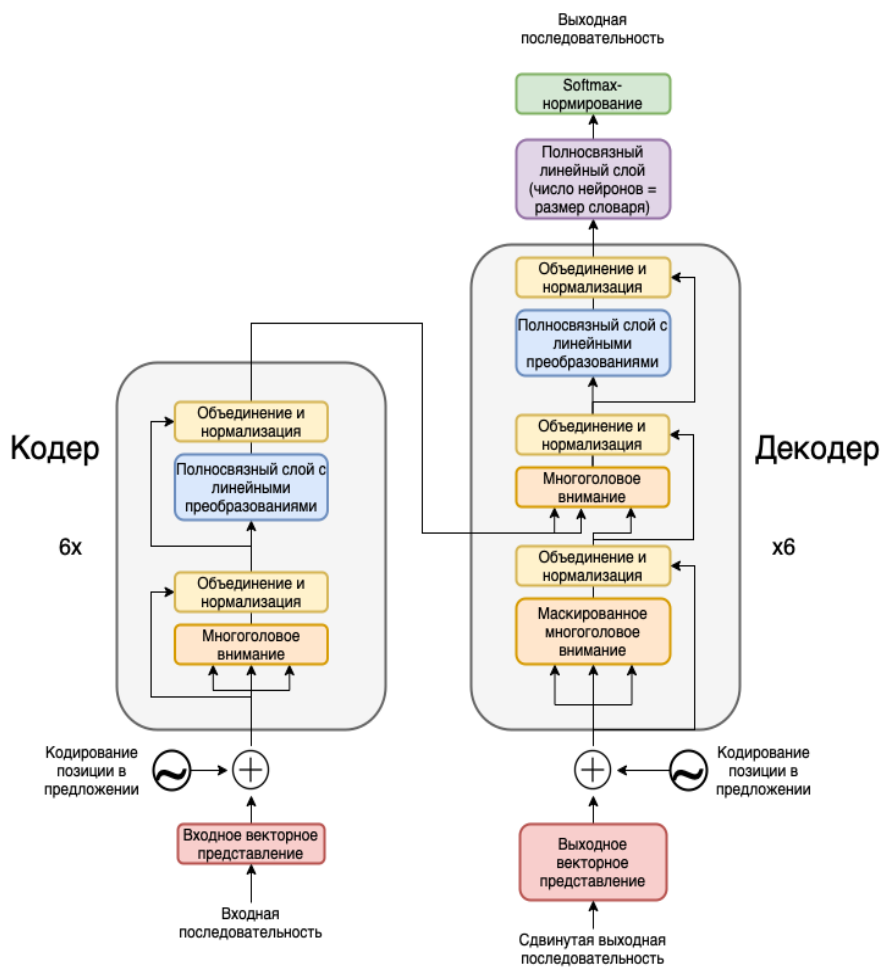


Рис. 2. – Схематичное изображение архитектуры трансформера

Реализация и настройка описанной архитектуры трансформера осуществлялась при помощи фреймворка «OpenNMT-ру» для языка программирования «Python» [9]. Данный фреймворк содержит большое количество инструментов для решения задач компьютерной лингвистики, обучения нейросетей машинного перевода и средств для их оценки.

Также важным преимуществом фреймворка «OpenNMT-ру» является то, что для ускорения процесса обучения нейронной сети можно использовать технологию параллельных вычислений «CUDA», что значительно сократило время обучения сети [10].

Обучение нейронной сети производилось с промежуточной оценкой параметров точности перевода (Accuracy) и растерянности (Perplexity). После достижения 30 эпох обучения динамика параметров уменьшилась, что говорит о необходимости прекращения процесса обучения нейронной сети во избежание негативного явления переобучения сети. Промежуточные параметры точности и растерянности представлены в таблице 1.

Таблица № 1

Промежуточные параметры нейронной сети

Эпоха	Accuracy	Perplexity
1	76.25	2.22
7	82.46	1.74
14	88.52	1.63
21	89.71	1.52
30	89.99	1.51

В итоге была реализована и обучена нейронная сеть, позволяющая переводить последовательности лемм в согласованные предложения, представленные на русском языке.

Заключение

Получившаяся после обучения нейронная сеть для перевода последовательностей лемм в связный текст была применена при разработке веб-приложения, с помощью которого можно создавать коллекции пиктограмм и составлять из них пиктограммные сообщения.

После составления пиктограммного сообщения клиентская часть веб-приложения отправляет на сервер, где развернута обученная нейронная сеть, последовательность текстовых интерпретаций каждой пиктограммы и получает в ответ согласованное предложение.

Применение приложения, осуществляющего перевод последовательности пиктограмм в связный русскоязычный текст, людьми с речевыми нарушениями, позволяет облегчить процесс коммуникации с теми, кто использует устную речь и не обладает навыками работы с пиктограммами. За счет этого разработанный метод перевода позволяет улучшить социальную адаптацию людей с речевыми нарушениями.

Литература

1. World Health Statistics // World Health Organization URL: who.int/gho/publications/world_health_statistics/2016/en/ (дата обращения: 10.12.2016).

2. Kultsova M.B., Matyushechkin D.S., Usov A.O., Karpova S.A., Petrenko A.A. Assistive technology for complex support of children rehabilitation with autism spectrum disorder // 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA). 2017. P. 1-5.

3. Смирнова И.А. Логопедическая диагностика, коррекция и профилактика нарушения речи у детей с ДЦП. Алалия, дизартрия, ОНР. 2004. 320 с.

4. Кульцова М.Б., Усов А.О., Матюшечкин Д.С., Карпова С.А. Мобильное приложение «Travel and communication assistant» для людей с

когнитивными нарушениями // Известия Волгоградского государственного технического университета. 2017. №8 (203). С. 64-71.

5. О машинном переводе // Технологии Яндекса URL: yandex.ru/dev/translate/doc/dg/concepts/how-works-machine-translation.html (дата обращения: 10.03.2021).

6. MyStem // Технологии Яндекса URL: yandex.ru/dev/mystem/ (дата обращения: 10.03.2021).

7. Матюшечкин, Д.С., Кульцова М.Б., Аникин А.В. Веб-сервис для перевода пиктограммных сообщений в согласованные тексты на русском языке // Известия Волгоградского государственного технического университета. 2018. №5(215). С. 30-36.

8. NLP Course // For You. URL: lena-voita.github.io/nlp_course/seq2seq_and_attention.html (дата обращения: 21.05.2022).

9. Шелудько В. М. Язык программирования высокого уровня Python. Функции, структуры данных, дополнительные модули: учебное пособие. 2017. 109 с.

10. OpenNMT-py Framework // OpenNMT-py URL: opennmt.net/OpenNMT-py/ (дата обращения: 05.02.2018).

References

1. World Health Statistics // World Health Organization URL: who.int/gho/publications/world_health_statistics/2016/en/ (Date accessed: 10.12.2016).

2. Kultsova M.B., Matyushechkin D.S., Usov A.O., Karpova S.A., Petrenko A.A. 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA). 2017. pp. 1-5.

3. Smirnova I.A. Logopedicheskaya diagnostika, korrekciya i profilaktika narusheniya rechi u detej s DCP. Alaliya, dizartriya, ONR [Speech therapy diagnostics, correction and prevention of speech disorders in children with cerebral palsy. Alalia, dysarthria, ONR]. 2004. 320 p.
4. Kul`czova M.B., Usov A.O., Matyushechkin D.S., Karpova S.A. Izvestiya Volgogradskogo gosudarstvennogo texnicheskogo universiteta. 2017. №8 (203). pp. 64-71.
5. O mashinnom perevode [About machine translation]. Texnologii Yandeksa URL: yandex.ru/dev/translate/doc/dg/concepts/how-works-machine-translation.html (Date access: 10.03.2021).
6. MyStem. Texnologii Yandeksa URL: yandex.ru/dev/mystem/ (Date access: 10.03.2021).
7. Matyushechkin, D.S., Kul`czova M.B., Anikin A.V. Izvestiya Volgogradskogo gosudarstvennogo texnicheskogo universiteta. 2018. №5(215). pp. 30-36.
8. NLP Course. For You URL: lena-voita.github.io/nlp_course/seq2seq_and_attention.html (Date access: 21.05.2022).
9. Shelud`ko V. M. Yazyk`k programirovaniya vy`sokogo urovnya Python. Funkcii, struktury` danny`x, dopolnitel`ny`e moduli: uchebnoe posobie posobie [The high-level programming language Python. Functions, Data Structures, Additional Modules: Tutorial]. 2017. 109 p.
10. OpenNMT-py Framework. OpenNMT-py URL: opennmt.net/OpenNMT-py/ (Date access: 05.02.2018).