А. О. Мельничук
г. Минск, Республика Беларусь

СПОСОБЫ И МЕТОДЫ ОБНАРУЖЕНИЯ ИСКУССТВЕННО СОЗДАННОГО ТЕКСТА

Генерация текста нейронными сетями стала доступна всем пользователям интернета в ноябре 2022 года с появлением ChatGPT 3.5. Это событие открыло новые возможности в сфере искусственного интеллекта и положило начало бурному развитию нейросетей. В тоже время, возникли и новые проблемы, с которыми ранее никто не сталкивался.

Проблема искусственной генерации затрагивает этическую и правовую сторону жизни. Вполне целесообразно использовать нейронные сети для создания рекламы в социальных сетях, но при написании квалификационной работы в университете прибегать к помощи подобных сервисов запрещено. Остро стоит вопрос безопасности: если нейросети все еще не очень хорошо справляются с написанием правдоподобного художественного текста, то новостные заметки получаются с трудом отличимыми от написанных человеком. В сочетании с хакерскими взломами, создание и распространение провокационных новостей может дезинформировать людей и в определенных ситуациях способствовать возникновению паники.

В связи с этими потенциальными угрозами выросло и отдельное направление, которое ставит своей целью распознавание в автоматическом режиме сгенерированных текстов. Такая задача представляется весьма непростой, учитывая быстрый прогресс в развитии нейросетей. Сгенерированные тексты уже сейчас трудно, а иногда невозможно отличить от написанных человеком. Однако при проведении более подробного исследования некоторые отличия найти все еще возможно.

Существует несколько способов выявить искусственно созданный текст. Основные отличия кроются в оформлении и содержании, однако некоторые характерные черты прослеживаются и на лингвистическом уровне: лексика и стилистика сгенерированных текстов обладают набором особых маркеров, по которым можно судить о характере происхождения текста. Если в пределах одного текста эти отличия для читателя незаметны, то при прочтении большого количества подобных генераций становится очевиден определенный шаблон, который лежит в основе всех искусственных текстов.

Следующие ниже результаты исследования были получены при анализе статей публицистического жанра в количестве 200 текстов на французском языке: 100 текстов сгенерированных и 100 текстов, созданных человеком. Для генерации текстов использовались нейросети ChatGPT и Cedille, а отбор настоящих статей осуществлялся в интернет-изданиях Le Figaro и Libération. Все статьи принадлежат к информационному жанру. Для более достоверных результатов исследования практически все сгенерированные статьи создавались на аналогичную тематику со статьями реальными, в некоторых случая заголовки французской прессы служили промптом (текстовым описанием для нейросети) для создания сгенерированного текста.

Первым заметным маркером становится разбиение текста на абзацы. Статьи французской прессы могут различаться по количеству и длине абзацев в зависимости от жанра и характера сообщения. Обычно это не менее двух абзацев, если речь идет о коротком информационном сообщении и до восьми абзацев, не считая подзаголовка, при более подробном освещении событий. При этом длина абзацев внутри одной статьи может сильно различаться либо же наоборот, примерно совпадать. Закономерность тут практически отсутствует. Сгенерированные ChatGPT тексты в 70 % случаев содержат по шесть абзацев, каждый по два предложения, что делает все абзацы в статье абсолютно одинакового размера.

Значительные различия можно также обнаружить на уровне лексического наполнения. Так, для сгенерированных текстов характерно частое повторение одного и того же слова или выражения на протяжении всего текста. При этом эти слова могут быть легко заменены синонимами, а выражения перефразированы для избежания тавтологии. Например, в одном из сгенерированных текстов слово *contribuer* ‘способствовать’ встречалось 5 раз, в другом – слово *violence* ‘жестокость’ употреблялось 4 раза, причем по два раза в одном абзаце, не считая употребления однокоренного слова *violent* ‘жестокий’. Другой текст изобиловал словом *escalade* ‘эскалация’ – 3 употребления и в том же тексте еще 2 употребления однокоренного слова *désescalade* ‘деэскалация’. Кроме прочего, слова *escalade* и *désescalade* ни разу не встречаются в анализируемых текстах французской прессы.

Не встречаются в сгенерированных текстах также аббревиатуры и сокращения, в то время как для настоящих французских статьей характерно их широкое употребление, что во многом затрудняет восприятие текста читателям, слабо знакомым с французскими реалиями. Среди сокращений часто встречаются названия партий (*AEI – Alliance écologiste indépendante* ‘Независимый экологический альянс’); названия предприятий (*SNCF – Société Nationale des Chemins de fer Français* ‘Национальное сообщество французских железных дорог’); названия медицинских процедур (*IVG – interruption volontaire de grossesse* ‘добровольное прерывание беременности’). Последняя аббревиатура, несмотря на широкое распространение во французских СМИ в связи с принятыми недавно изменениями в Конституции Франции, ни разу не была сгенерирована нейросетями.

По естественным причинам нейросети также избегают упоминания имен собственных, в особенности имен людей, и цитирования. Для настоящих СМИ характерно широко ссылаться на источники информации, в первую очередь на должностных лиц, от которых поступили официальные сведения, и прямо или косвенно цитировать их слова [1, с. 69]. Кроме того, французские газеты часто прибегают к *intercitation*, что можно перевести на русский язык как ‘взаимное цитирование’: явление, когда одна газета ссылается на другую. Нейросети не способны создать такую сложную систему указания источников информации, поэтому нет упоминания ее происхождения.

Особый интерес для лингвистов представляет собой список слов, которые нейросети по неопределенным причинам употребляют значительно чаще, чем журналисты, и наоборот. При проведении данного исследования был создан корпус, в котором было проведено статистическое исследование. Ниже в таблице приведен список некоторых таких слов.

Таблица

Частотность употребления некоторых слов
в сгенерированных текстах и текстах СМИ

Слово	Количество словоупотреблений на 100 единиц текста			
	Сгенерированные тексты			Тексты СМИ
	ChatGPT	Cedille	Всего	
Crucial ‘жизненно важный’	25	9	**36**	0
Face à ‘перед лицом’	17	21	**38**	6
Tensions ‘напряжение’	13	23	**36**	2
Impératif ‘обязательный’	13	7	**20**	0
Parquet ‘прокуратура’	0	0	**0**	34
Cependant ‘однако’	16	14	**30**	1
Violence ‘жестокость’	18	19	**37**	1
Droits ‘права’	26	24	**50**	14
Sécurité ‘безопасность’	74	46	**120**	9
Région ‘регион’	46	17	**63**	9
Selon ‘согласно’	9	6	**15**	43

Список выявленных слов, сильно различающихся по частоте употребления, намного шире. Так, в сгенерированных текстах часто употребляются вводные слова и слова-связки, практически полностью отсутствующие в текстах СМИ. Из-за отсутствия ссылок на источники реже употребляются выражения типа *selon* ‘согласно’. Некоторые выражения, встречающиеся в искусственных текстах, отличаются по своей форме от аналогичных выражений в текстах, созданных человеком. Так, сгенерированные тексты предпочитают выражение *dans ce contexte* ‘в данном контексте’, в то время как в реальных текстах употребляется несколько иная конструкция *dans un contexte*; в сгенерированных текстах употребляется слово *vie* ‘жизнь’ во множественном числе, в текстах СМИ только в единственном; в искусственных текстах встречается редко употребляемое слово *différend* ‘спор’, которое не используется в публицистическом жанре.

Некоторые слишком частые по сравнению с реальными текстами словоупотребления также напрямую связаны со стилистикой сгенерированных текстов.

Публицистический стиль широко использует различные экспрессивные средства выражения, такие как качественно-оценочные прилагательные и существительные, некоторые фразеологизмы, которые нередко становятся

газетными штампами [2, с. 79]. В тоже время солидные газетные издания избегают сгущения красок и нагнетания ситуации, отдавая предпочтение сдержанным и спокойным выражениям.

Сгенерированные тексты, в свою очередь, используют больше эмоционально-оценочной лексики, которая легко воздействует на читателя. В особенности это касается прилагательных, несущих негативную оценку, таких как *imminent* ‘неминуемый’, *dévastateur* ‘разрушительный’, *aveugle* ‘слепой’ (в контексте *слепая ярость*, *слепая ненависть*), *effrayant* ‘ужасающий’. Также нередко употребление схожих по эмоциональному воздействию существительных, например: *craintes* ‘страхи’, *souffrances* ‘страдания’, *méfiance* ‘настороженность’. В употреблении подобных слов важна не только их частота употребления относительно прессы, но и насыщенность каждого сгенерированного текста подобными словами и выражениями, которая превышает подобные показатели реальных текстов СМИ.

На уровне грамматики также имеются некоторые подсказки, способные указать на происхождение текста. Так, в сгенерированных текстах в три раза реже встречается употребление оборота Conditionnel Présent (условное наклонение), которое в публицистическом дискурсе используется для указания на непроверенные факты, которые ждут подтверждения. В переводе на русский такие предложения переводятся с использованием конструкций *по поступающим сообщениям*, *ожидается*, *предполагается*.

Что касается смыслового содержания, то и здесь отличия достаточно значительны. Например, большинство статей французской прессы, относящихся к жанру *faits divers* ‘происшествия’ заканчиваются либо статистикой подобных происшествий, либо кратким описанием похожих случаев со ссылками на соответствующие статьи. Статьи жанра *actualités internationales* ‘международные новости’ завершаются кратким обзором цепочки предыдущих событий, которые привели к текущему положению дел, либо же цитированием той или иной персоны. Метеорологические сводки – историческими рекордами погодных показателей. При генерировании текста зачастую последним абзацем становится подведение итогов всей статьи, активно призывающее к тем или иным действия: начать мирные переговоры, срочно приступить к борьбе за права женщин или принять меры против глобального потепления. В результате намного чаще употребляются такие вводные фразы как *en conclusion* ‘в заключение’, *pour conclure* ‘в заключение’, *comme résultat* ‘как результат’, а также слова, выражающие безотлагательность или обязательность действия, такие как *impératif* ‘обязательный’ или *urgent* ‘срочный’.

Что касается существующих методов распознавания сгенерированных текстов, то на данный момент их насчитывается три вида [3, с. 9963]:

1. Основанные на использовании лингвистических особенностей, направленных на различение уникальных стилей письма нейронных сетей и людей. Из-за вычислительных затрат на извлечение лингвистических данных исследователи предложили детектор, основанный на глубоком обучении, т.е. нейронную сеть, которая будет способна опознавать тексты, написанные другими нейронными сетями. Однако они оказались легко вос-

приимчивы к так называемым вредоносным возмущениям – специальным образом сформированным данным, которые подаются на вход нейронной сети и сбивают ее с толку. К тому же, им требуется большой датасет, т.е. объем данных для изучения, для получения качественного результата.

2. Основанные на статистических данных. Они не поддаются враждебным возмущениям и требуют минимального количества данных. Однако такой способ уступает в производительности первому.

3. Основанные на сочетании лингвистических и статистических методах, чтобы добиться устойчивости и высокой производительности.

Таким образом, разработка таких алгоритмов помогает в автоматическом режиме отслеживать маркеры, указывающие на искусственную генерацию текста. Разумеется, их результаты не могут быть абсолютно точными. Однако, когда-нибудь будет возможно включить функцию автоматического распознавания, которая будет анализировать тексты, которые пользователь читает на электронных устройствах и активировать предупреждения о том, что были обнаружены подозрительные признаки его искусственного происхождения.

Нейронные сети развиваются с поразительной скоростью. При подготовке к написанию этой статьи выяснилось, что ChatGPT при генерации статьи указывает дату и место «публикации» без дополнительного запроса от пользователя. Еще в начале марта этого года такие возможности не наблюдались. Такие темпы развития подтверждают актуальность исследований в сфере распознавания искусственно созданных текстов.

ЛИТЕРАТУРА

1. Особенности языка и стиля современной французской печати : сб. науч. ст. / Вестник РУДН ; редкол.: А. Г. Коваленко (гл. ред.) [и др.]. – М. : РУДН, 2010. – 200 с.

2. Французская пресса через призму выразительных средств языкового пространства : сб. науч. ст. / Научный журнал КубГАУ ; редкол.: А. И. Трубилин (гл. ред.) [и др.]. – Краснодар : КубГАУ, 2015. – 210 с.

3. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark : Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, December 2023 ; ed.: H. Bouamor [et al.]. – Oxford, Miss. : Univ. of Mississippi, 2023. – 9987 p.