А. О. Мельничук

АВТОМАТИЗИРОВАННЫЕ СПОСОБЫ РАСПОЗНАВАНИЯ ИСКУССТВЕННО СОЗДАННОГО ТЕКСТА

Искусственные тексты – это такие тексты, которые были написаны нейронными сетями без участия человека. В противоположность им выделяют естественные тексты. Существует несколько способов автоматического отнесения текста к одной из данных категорий.

Метод стилометрии (stylometry, от слов «style» и «measurement») используется не только для детекции сгенерированных текстов, но и для установления их авторства. Метод стилометрии основан на автоматическом анализе таких характеристик как синтаксис и выбор лексики, которые позволяют определить автора текста (Helena Gomez-Adorno, 2018). При обнаружении сгенерированного текста стилометрия может быть использована для определения отличительного стиля конкретной нейронной сети на основе определенных особенностей, таких как словоупотребление, структура предложений и других лингвистических закономерностей.

Следующий способ основан на глубоком обучении, т.е. представляет из себя нейронную сеть, которая будет способна опознавать тексты, написанные другими нейронными сетями. Для этого им требуется объемный размеченный датасет из текстов, написанных человеком, и текстов, написанных нейросетями. В процессе обучения они выявляют существующие различия, благодаря чему могут впоследствии быстро делать достаточно точные заключения.

Существуют также методы, основанные на статистических данных. Обычно они используют классические методы машинного обучения, такие как линейная регрессия, метод опорных векторов (SVM), наивный Байесовский классификатор и другие. Они требуют явного определения признаков и обучаются на основе статистических закономерностей в данных.

Самыми эффективными на данный момент являются методы, сочетающие методы глубокого обучения и статистические методы, чтобы добиться устойчивости и высокой производительности. Это достигается путем кодирования статистических признаков во входных данных. Такой специальный слой, добавленный к обученной нейронной сети, повышает ее устойчивость.

Как мы видим, на данный момент не существует единого универсального способа детекции искусственных текстов. Тем не менее, их комбинация позволяет с относительно высокой точностью делать выводы о происхождении текста.