создавались алфавитно-частотные словари. Для последующего выделения из такого словаря опорных слов будущих терминологических словосочетаний с каждым словом полученного частотного словаря делались следующие преобразования:

1. Удалялись служебные и общеупотребительные словоформы. Эти словоформы часто называют «запрещенными» словами. Например, из словарей были исключены артикли (в английских словарях), предлоги, союзы, частицы, вспомогательные глаголы, числительные, общеупотребительные словоформы и имена собственные.

2. Суммирование частоты употребления всех грамматических форм одного и того же слова. Например, слово *image* встретилось в формах единственного и множественного числа:

IMAGES 7 6 6:1 15:1 17:1 18:1 20:1 22:2
IMAGE 1 1 20:1

После объединения грамматических форм в словаре остается одна словоформа image с обще частотой употребления (F) 8(F=7+1=8) и общим количеством абзацев (n) 7 с учетом их повторяемости.

Также были удалены словоформы, которые употреблены только в одном абзаце, поскольку они относятся к содержанию только этого абзаца, а не всего текста.

Третий этап заключался в непосредственном выделении списков главных опорных слов текста. Для этого для каждого слова выделялся коэффициент семантической значимости слова (КВ), который вычисляется по формуле:

$$КВ=\frac{F*m}{L*n},$$

где F – частота слова в тексте с учетом проведенных преобразований; m – число абзацев, в которых употреблено слово; L – общее количество слов в тексте; n – общее количество абзацев в тексте.

Для последующего отбора опорных слов будущих терминологических словосочетаний полученные слова текста сравнивались со списком опорных слов по вычислительной технике.

В. Лускин

РАСПОЗНАВАНИЕ ТЕКСТА С ПОМОЩЬЮ НЕЙРОННЫХ СЕТЕЙ

Одной из важных задач нашего времени является сохранение информации.

Наиболее простым способом ввода информации с бумажного носителя в память компьютера является сканирование документа. Результатом сканирования является графический файл, который затем преобразуют в текстовую форму. Процесс такого преобразования называют процессом распознавания. Существует несколько основных методов распознавания текста:

1) сравнение с заранее подготовленными шаблонами;
2) распознавание с использованием критериев распознаваемого объекта;

3) распознавание при помощи самообучающихся алгоритмов, в том числе нейронных сетей.

В настоящее время большую популярность набирает метод распознавания текстов с помощью нейронных сетей. Важным является то, что нейронная сеть – это обучаемая система. Она действует не только в соответствии с заданным алгоритмом и формулами, но и на основании прошлого опыта. Нейронная сеть – это математическая модель, чей принцип работы схож с биологическими нейронными сетями – сетями нервных клеток живого организма.

Как правило, нейронные сети состоят из нескольких слоев, элементы которых называют нейронами. Каждый нейрон представляет собой процессор, выполняющий обработку поступившей информации. Выделяют входные нейроны, выходные, и скрытые.

На вход системы распознавания поступает растровое изображение документа. Производится сегментация текста на структурные единицы: строки, слова и отдельные символы. Сегментация на строки и слова осуществляется на основе расстояний между темными областями в изображении.

Невозможно добиться высоких результатов без предварительного обучения нейронной сети. Для ее обучения создаются специальные выборки объемом в десятки тысяч случаев. Нейронная сеть получает на вход обучающую подборку и правильные ответы. Стоит отметить, что нельзя гарантировать безошибочное распознавание символов в силу различного их начертания в зависимости от шрифтов, а также возможных дефектов начального изображения. Тем не менее, процент правильно распознанных символов зачастую находится в районе 97–98 % для обучающей подборки.

Традиционные алгоритмы распознавания иногда не способны корректно распознать текст. Например, при низком качестве графического файла в процессе распознавании текста, написанного на французском языке, часто возникают ошибки, связанные с наличием во французском языке диакритических знаков. Части таких знаков могут быть восприняты системой распознавания как дефект изображения. Благодаря своим способностям к обучению наиболее эффективным является использование нейронных сетей для распознавания символов со сложным написанием, будь то диакритические знаки, иероглифическое письмо, или рукописный текст.

М. Марковская

ЯЗЫКОВЫЕ НОВАЦИИ СОЦИАЛЬНОЙ СЕТИ INSTAGRAM

Коммуникация в социальных сетях, будучи относительно новым явлением, является недостаточно исследованной с лингвистической точки зрения. Во-первых, это связано с тем, что интернет-коммуникация представляет собой крайне динамичную сферу функционирования языка. Во-вторых, под влиянием глобализации язык стал особенно восприимчивым к изменениям