5) корпусы параллельных текстов помогают разрешить проблему многозначности терминов при переводе, а также позволяют при наличии синонимичных вариантов переводов найти наиболее распространенный среди специалистов термин;

6) технология переводческой памяти избавляет переводчика от необходимости повторно переводить идентичные фрагменты текста, входящие в разные документы;

7) системы с переводческой памятью дают возможность установления единства терминологии при работе нескольких переводчиков над текстами определенной тематики.

Несмотря на все положительные характеристики рассматриваемого подхода к автоматическому переводу, существует ряд затруднений при разработке и применении систем с переводческой памятью.

Одной из основных проблем таких систем машинного перевода является параллельная сегментация, или установление соответствий между текстом оригинала и его переводом или переводами на уровне слов, предложений, абзацев или целых текстов. Стыковка текстов может осуществляться вручную или автоматически. Специалисты, занимающиеся проблемами разработки программного обеспечения для решения этой задачи, пока еще делают первые шаги в этой области и опираются, в основном, на структурное сходство оригинального текста и его перевода. Специалисты выделяют три подхода к автоматическому или полуавтоматическому выравниванию параллельных текстов: лингвистический, статистико-вероятностный и параллельная сегментация исходного текста и текста перевода.

Среди прочих проблем использования технологии переводческой памяти можно выделить необходимость постоянного пополнения баз данных достаточно большими массивами параллельных текстов разной тематики, а также автоматическое распространение неточностей, при их наличии, на весь проект.

Все вышеперечисленное дает нам право утверждать, что использование технологии переводческой памяти при автоматическом переводе научно-технических текстов способствует повышению качества машинного перевода, основанного на правилах, за счет, в первую очередь, точности перевода терминологической лексики, при общем сокращении времени, необходимого на выполнение перевода, и уменьшении объема работы переводчика.

М. В. Чернышевич

АВТОМАТИЧЕСКОЕ РАСПОЗНАВАНИЕ НЕДОСТОВЕРНЫХ МНЕНИЙ В СООБЩЕНИЯХ ПОЛЬЗОВАТЕЛЕЙ НА АНГЛИЙСКОМ ЯЗЫКЕ

В связи со стремительным ростом популярности социальных платформ для общения пользователей коммерческие компании и публичные деятели стали уделять большое внимание присутствию собственных названий и имен в социальных сетях. По данным компании Altimeter в 2016 г. 38 % коммерческих компаний инвестировали в рекламу в социальных сетях более

20 % всего бюджета на рекламу. Огромное количество пользователей, большие возможности инвестирования, а также отсутствие проверки достоверности сообщения или отзыва сделали социальные сети и сайты отзывов целью для мошеннических атак. Так, из 4840 профилей, созданных от имени известных компаний, 19 % являются «поддельными». Их создатели стремятся заработать на популярности других брендов или повлиять на мнение пользователей о товарах, услугах и др. Кроме того, спам (его подробное определение дается ниже) – это дешевый способ размещения в социальных сетях ссылок для поднятия рейтинга (PageRank) сайта в поисковых системах.

Спам-сообщения, которые в большом количестве присутствуют на целевых интернет-ресурсах и являются крайне нежелательными для обработки системой автоматического сентимент-анализа текста (АСАТ), можно разделить на следующие группы:

тип 1 (недостоверные отзывы) – отзывы, в которых автор намеренно дает заведомо ложную информацию, хвалебную или критическую, с целью обмануть читателя или систему автоматического сентимент-анализа;

тип 2 (реклама) – рекламные сообщения, которые чаще всего описывают бренд, а не отдельные товары и услуги, и являются результатом автоматической генерации. Например, *@Sephora: Amazing Clinique Acne Solutions Healing Gel. Use code SPOTGEL @ for more info visit our app http://t.co/cL0kPG*;

тип 3 (контент низкого качества) – к данному типу мы будет относить сообщения, в которых присутствует большое количество «мусорных» слов, не закрепленных в словаре анализируемого ЕЯ (и в списках сленговых слов) и затрудняющих понимание всего высказывания, например, *02 acura tl### Fully loaded## Sunof!!! http://u.co/PfGjsh*.

Все эти типы сообщений мы будем далее объединять названием *спам-сообщения*.

Так как сентимент-анализ предполагает агрегацию субъективных мнений реальных пользователей, то фильтрация спам-сообщений является важным этапом системы АСАТ. Анализ показал, что сообщения типа 2 и 3 легко могут быть идентифицированы экспертом и размечены в корпусе для последующего обучения классификатора разрабатываемой системы. Сложнее дело обстоит с сообщениями типа 1, так как они зачастую написаны реальными людьми и их сложно отличить от настоящих отзывов без анализа метаинформации (внетекстовой информации), такой как, например, количество отзывов об одном и том же бренде (продукте) от одного и того же пользователя или дубликаты отзывов от разных пользователей. Сбор и анализ подобной метаинформации значительно усложняет архитектуру системы АСАТ. Кроме того, следует учитывать, что частота сообщений типа 1, как оказалось, значительно меньше, чем сообщений других типов, поэтому целевая система АСАТ должна в первую очередь фильтровать сообщения типов 2 и 3.

Задача идентификации спама решается как задача классификации с использованием методов машинного обучения с учителем, а это требует наличия обучающего корпуса, состоящего из сообщений пользователей, аннотированных метками класса «спам» и «не спам». Несмотря на то, что, спам-сообщения довольно легко распознать эксперту, разметка корпуса,

учитывая, что он должен быть достаточно большого размера, является очень трудоемким процессом, и его желательно автоматизировать. С этой целью было выбрано 7 178 685 произвольных сообщений. Для автоматического формирования из них выборки сообщений, принадлежащих классу «спам», были сформулированы на основании проведенного анализа следующие критерии.

- *Повторяемость в выборке более 20 раз (полные или частичные дубликаты):* данный критерий позволяет выбрать автоматически сообщения, которые чаще всего соответствуют спам-сообщениям типа 2. Например, в исходной выборке 1 590 раз встретилось сообщение *check this out! I made $300 today so far URL* 'Зацени-ка это! Я заработал сегодня 300$ URL'.
- *Отношение «мусорных» слов сообщения к общему количеству его слов более чем 0.3:* данный критерий позволяет выделить спам-сообщения типа 3.
- *Наличие устойчивых конструкций, указывающих на рекламный характер сообщений и, таким образом, на их соответствие спам-сообщениям типа 2:* это конструкции типа

  available online . URL* 'доступен онлайн .* URL',
  special price . URL* 'специальная цена .* URL',
  visit our website . URL* 'зайдите на наш сайт .* URL',
  contact us . URL* 'свяжитесь с нами .* URL' и др.

  Всего разработано 26 паттернов, ориентированных на данный критерий.

Отметим, что сообщение помещалось автоматически в формируемую выборку, если оно удовлетворяло хотя бы одному из представленных критериев. Для формирования выборки сообщений, относящихся к классу «не спам», из множества сообщений пользователей, не попавших в первую выборку, было взято произвольным образом 8 000 уникальных сообщений, т.е. повторяемость которых равна 1. Кроме того, из оставшегося множества вручную было размечено еще 10 000 сообщений пользователей, одна часть которых попала в первую выборку, а другая – во вторую. Таким образом, было выбрано всего 25 583 сообщений (10 214 для класса «спам» и 15 369 для класса «не спам»).

Имея обучающую выборку, система, основанная на методах машинного обучения с учителем, далее автоматически создает для себя набор правил, используемых ею для обучения классифицирующего алгоритма или просто классификатора, т.е. в конечном счете, для решения целевой задачи, в данном случае – распознавания спам-сообщений в неаннотированных сообщениях. При этом ключевую роль играет набор признаков, которыми оперирует классификатор, а формирование и обоснование этого набора – функция эксперта, и она реализуется путем тщательного анализа обучающей выборки.

В качестве классифицирующего алгоритма в данной работе выбран метод опорных векторов (SVM), так как он показал наилучшее качество и наибольшую скорость обработки выборки. Для обучения классификатора было исследовано большое количество признаков, предложенных в различных работах. В конечном счете сформированное пространство признаков включило в себя такие лексические, стилистические, синтаксические и статистические признаки, как:

- бинарные признаки наличия или отсутствия униграмм и биграмм в сообщении;

- количество хэш-тэгов в сообщении;
- количество ссылок в сообщении;
- количество смайлов в сообщении;
- количество слов в сообщении;
- отношение количества слов сообщения, принадлежащих словарю ЕЯ, к общему количеству слов сообщения;
- количество САО-отношений в сообщении;
- количество личных местоимений в сообщении;
- отношение количества слов сообщений, написанных с заглавной буквы, по отношению к общему количеству в нем слов;
- отношение количества слов сообщения, написанных с маленькой буквы, по отношению к общему количеству в нем слов.

Тестирование разработанного модуля фильтрации спам-сообщений показало следующие оценки эффективности его работы на контрольной выборке: точность 98,76 % и полнота 97,68 %.

Анализ результатов обработки данным модулем пользовательских сообщений показал, что спамом являются около 35 % сообщений из социальной сети Twitter (714 из 2059), около 20 % сообщений из сети Facebook (382 из 1986) и около 2 % – из форумов Fixya и Amazon (45 из 2533).

Н. Г. Швец

ВЗАИМОСВЯЗЬ ВЕРБАЛЬНЫХ И ВИЗУАЛЬНЫХ СОСТАВЛЯЮЩИХ КРЕОЛИЗОВАННОГО ТЕКСТА

Стремительное развитие новых информационных технологий значительно расширило круг интересов лингвистики, выдвинуло в качестве одной из первоочередных задач всестороннее исследование языка во всем многообразии его связей с кодами других семиотических систем.

В последние годы в лингвистике текста объектом лингвистического исследования становятся так называемые креолизованные тексты, т.е. тексты, фактура которых состоит из двух частей: вербальной и невербальной. Обращение к невербальным средствам связано с поиском закономерностей текстообразования, выявлением роли этих средств в организации прагматического воздействия текста на адресата, необходимостью более полного извлечения текстовой информации.

Мир креолизованных текстов чрезвычайно многообразен. Он охватывает тексты газетно-публицистические, научно- технические, тексты-инструкции, иллюстрированные художественные тексты, афиши, комиксы, карикатуру, шарж, плакаты, открытки, географические карты и др. К креолизованным текстам относятся и печатные рекламные объявления, содержащие иллюстрацию.

Динамика взаимоотношений вербальных и визуальных элементов в рекламном объявлении весьма разнообразна и определяется видом рекламы. Американский исследователь Р. Бартес выделил два наиболее часто встречающихся типа отношений текста и изображения: *якорь* и *реле*. При первом типе