сноски, ремарки, цитаты, таблицы, формулы). При анализе текстов учитывается как языковое, так и структурное членение текста. Многие структурные элементы текста имеют яркие языковые особенности, которые требуют специальной формы представления. Разметка – это расстановка в тексте документа специальных маркеров (тегов), которые эксплицируют «скрытые» элементы информации, присутствующие в тексте. Разметка широко используется в корпусах текстов, электронных словарях и т.д. Электронный словарь подразумевает определенный способ выборки словника, множество текстов, из которых делается эта выборка, дополнительную информацию, которая сопровождает каждое слово.

В настоящее время лексикография находится под сильным воздействием новых методов обработки информации. Структура словарной статьи в разных электронных словарях различна. Словарная статья содержит разнотипную лингвистическую информацию. Словари включают в свои словарные статьи следующие сведения: заглавное слово, его тематическую принадлежность, грамматическую информацию, неформализованные толкования или стандартизированные дефиниции, лексическую сочетаемость, семантические иерархические связи заглавного слова, стилистическую окрашенность, лингвогеографические ограничения, контексты, различную служебную информацию.

В нашей работе материалом исследования служат англоязычные тексты по экономике, взятые с Интернет–сайта http://marketwatch.com. По всем анализируемым текстам были получены алфавитно-частотные словари с помощью программы DICT. Практическое теггирование полученных словарей по проанализированным англоязычным текстам осуществлялось на основе системы кодирования, разработанной на кафедре информатики и прикладной лингвистики. Так, например, согласно кодировочной таблице существительное *economy* кодируется следующим образом: *ECONOMY* – NO1S42; *GLOBAL* – AQPOOO21. Нами были закодированы все тексты из отобранного массива англоязычных текстов.

Для пополнения лингвистической информационной базы данных нами были использованы программы, которые позволяют накладывать закодированный словарь на текст. Для получения закодированного текста между словом и тегом был проставлен символ «_» без пробела, каждое слово вводилось с новой строки. Тем самым, каждое словоупотребление текста получает соответствующую информацию о различных лингвистических признаках.

В. Керножицкая

ПОДХОДЫ К АВТОМАТИЧЕСКОЙ НОРМАЛИЗАЦИИ ЛЕКСИЧЕСКИХ ЕДИНИЦ В СООБЩЕНИЯХ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ

Участники сетевого общения зачастую используют формы выражения мыслей, далекие от традиционных лексических норм. Так, в текстах интернет-коммуникации встречаются аббревиатуры, отражающие, по сути,

реакцию пользователя (*LOL – laughing out loud*), сокращения (*abt – about*), намеренные орфографические ошибки, опечатки, повторяющиеся символы и заглавные буквы (*OOOO!*), нестандартное использование знаков пунктуации и пробелов, такие графические символы как эмотиконы (смайлы) и эмодзи. С лингвистической точки зрения лексическая форма их представления далека от традиционной, принятой в конкретном языке, что делает практически невозможным их автоматический анализ до тех пор, пока они не будут нормализованы.

Рассмотрим подробнее основные подходы к автоматической нормализации лексических единиц на примере обработки сообщений англоговорящих пользователей социальной сети Twitter. Так, разработан метод нормализации твитов, совмещающий в себе как методы машинного обучения, так и правила. На начальном этапе происходит отбор всех некорректных слов с помощью модели машинного обучения, натренированной на их поиск по определенным атрибутам. Затем к каждому слову применяется одно из следующих правил нормализации: 1) расшифровка аббревиатур (при этом используется список самых частотных аббревиатур в сети Twitter и их расшифровок); 2) обработка причастий настоящего времени (пользователи часто пропускают в них букву *i* или *g*, например, употребляют *goin* вместо *going*. Данное правило направлено на устранение таких опечаток: сначала определяется часть речи, и если это причастие, то восстанавливается его нормальная форма); 3) обработка пропущенного апострофа (правило направлено на исправление ошибок в таких случаях, как *'m, 'll, 've, 're, n't, 's*); 4) обработка повторяющихся символов (в данном случае происходит удаление избыточного количества повторяющихся символов). Группа исследователей представила еще один подход под названием *TENOR* (*TExt NORmalisation Approach*). Процесс нормализации, предложенный ими, состоит из двух этапов – отбора слов, не соответствующих нормам, с помощью словаря и их дальнейшей замены на стандартную форму. Процесс замены включает в себя следующие действия: все нестандартные символы и пунктуация, за исключением смайлов, удаляются с помощью регулярных выражений, затем расшифровываются аббревиатуры, смайлы, сокращения и устраняются опечатки. Разработана еще одна система лексической нормализации, состоящая из двух блоков. В первом блоке происходит выделение потенциальных кандидатов для нормализации с использованием метода *CRF* (*Conditional Random Fields*, разновидность метода марковских случайных полей), натренированного на корпусе из 2950 размеченных твитов, а во втором блоке – их последующая нормализация. Стратегия лексической нормализации твитов, предложенная еще одной группой разработчиков, включает в себя следующие шаги: 1) генерацию набора слов, являющихся потенциальными кандидатами на нормализацию; 2) определение всех возможных вариантов для нормализации каждого слова, отбор наиболее вероятного варианта путем сравнения с помощью алгоритмов, которые позволяют, например, в зависимости от контекста привести слово *mve* к слову *me* или к слову *move*; 3) приведение слов к их стандартной форме.

Таким образом, в общем плане процедура автоматической нормализации лексических единиц осуществляется в два этапа. На первом этапе происходит отбор не соответствующих лексическим нормам слов твита, а на втором этапе – их последующая нормализация либо на основе стандартных алгоритмов, либо методами машинного обучения.

Е. Кореба

ЯЗЫКОВЫЕ МАРКЕРЫ ОТРАЖЕНИЯ РЕПУТАЦИИ ЛИЧНОСТИ В АНГЛОЯЗЫЧНЫХ ТВИТАХ

Репутация личности – это устоявшееся общественное мнение о конкретном человеческом индивиде как субъекте отношений и сознательной деятельности, которое является результатом оценивания прошлого поведения индивида группой людей и используется с целью прогнозирования его дальнейшего поведения. В зависимости от основных характеристик конкретного индивида его репутация имеет определенную структуру. Так, структура репутации политического лидера включает в себя такие основные психосемантические факторы, как предназначение власти (отношение лидера к своей стране и народу); сила личности (волевая регуляция поведения, решительность, принципиальность, трудолюбие и интеллект); нравственно-этическая позиция лидера (моральные и душевные качества и ценности политика). Дополнительными факторами являются выразительность самопрезентации, стиль руководства, коммуникативная установка, религиозность, тип мышления и мускулинность – феминность.

Материалом исследования послужили тексты 500 англоязычных твитов, в которых высказываются различные мнения о Президенте США Д. Трампе. Тексты твитов публиковались с января 2017 по апрель 2019 года. Особенностью отобранных твитов является то, что представленная в них информация носит, преимущественно, субъективный характер, т.е. содержит оценочную и эмоциональную составляющие. В ходе анализа отобранного массива твитов были выявлены языковые маркеры отражения репутации Д. Трампа. Прежде всего, к ним относится оценочная и эмоционально-коннотативная лексика, например, *strong, appreciate, stupid, ignorant, racist*: *Trump's too stupid to understand policy. He's an impotent and ignorant president weakening our country.* В ходе анализа твитов было выяснено, что хэштеги также могут нести в себе определенную оценку и напрямую влиять на общую репутацию Президента США. Например, *#MAGA* (расшифровывается как *Вернем Америке былое величие*) используется сторонниками Д. Трампа, а большинство его противников употребляют *#Resist* или *#Impeach*: *Replying to @realDonaldTrump Look at all that American Patriot support!!!! It's all for you, President Trump, because you are doing the right thing, sir!!!! #MAGA*; *When you build an administration based on lies, corruption and incompetence, you end up with @realDonaldTrump #TrumpRussia #Resistance.* Анализ материала исследования показал, что авторы твитов применяют различные интенсифи-