Евросоюза. В данной связи важно, как концепт «беженец» (*Flüchtling*) представлен в немецком лингвистическом корпусе. Словарь Ожегова дает следующее определение: «человек, оставивший место своего жительства вследствие какого-нибудь бедствия».

Для исследования речевой специфики актуальна информация, в частности «Мангеймского корпуса немецкого языка» (*COSMAS corpora* или *DeReKo*), содержащего *44,5* млрд словоупотреблений и крупнейшую коллекцию немецких текстов. Для поиска был использован подкорпус письменной речи с общим количеством *1 370 766 704* словоупотреблений. Так, первое упоминание лексемы *Flüchtlinge* зафиксировано в 1787 г. в произведении Гете «Страдания юного Вертера»:

Leute von einigem Stande werden sich immer in kalter Entfernung vom gemeinen Volke halten, als glaubten sie durch Annäherung zu verlieren; und dann gibt's ***Flüchtlinge*** *und üble Spaßvögel, die sich herabzulassen scheinen, um ihren übermut dem armen Volke desto empfindlicher zu machen.*

При этом для поиска были заданы следующие словоформы *беженец*: *Flüchtling*, *Flüchtlinge*, *Flüchtlings*, *Flüchtlingen*.

На запрос *Flüchtling* было найдено *44 784* словоупотребления; наибольшее число словоупотреблений (*8 295*) приходится на 2016 г. На запрос *Flüchtlinge* обнаружено *526 373* словоупотребления, – наибольшее число словоупотреблений (*124 114*) приходится на 2015 г. На запрос *Flüchtlings* обнаружено *8 782* словоупотребления, наибольшее число словоупотреблений (*1 348*) приходится на 2016 г. На запрос Flüchtlingen – обнаружено *151 058* словоупотреблений; наибольшее число словоупотреблений (*35 559*) приходится на 2016 г.

В ходе работы с корпусом было зафиксировано *730 997* словоупотреблений концепта «беженец» (*Flüchtling*), что подтверждает актуальность обсуждаемой темы в немецкой речи. Наибольшая частотность словоупотреблений приходится на период 2015–2016 гг. Эта тенденция очевидным образом подтверждает концептуальную актуальность «европейского миграционного кризиса». Таким образом, события, происходящие в стране, отражаются не только на социальной и политической сфере, но и влияют на языковую картину мира немецкого народа.

А. Скворцова

ПРИНЦИПЫ ФОРМИРОВАНИЯ БАЗЫ ДАННЫХ СИСТЕМЫ АВТОМАТИЧЕСКОГО АНАЛИЗА ОБЩЕСТВЕННОГО МНЕНИЯ ОБ ОРГАНИЗАЦИИ

Одной из самых востребованных задач в области автоматической обработки больших текстовых массивов является задача извлечения из них разного рода информации. Хорошим источником для ее решения являются тексты отзывов, содержащие авторские мнения (оценки) о некотором объекте

действительности. С целью обмена информацией между пользователями подобные тексты можно найти на различных площадках сети Интернет. Для извлечения мнений пользователей Сети используются два основных подхода. Первый подход называется инженерным, поскольку основан на заранее созданных словарях и правилах извлечения объектов. Второй подход использует методы машинного обучения.

В докладе рассматриваются принципы формирования базы данных системы автоматического анализа общественного мнения об организации на основе инженерного подхода. Материалом исследования послужили более 500 текстов англоязычных отзывов о предприятиях гостиничного и ресторанного бизнеса Республики Беларусь, взятых со специализированных сайтов отзывов http://www.tripadvisor.com и http://www.booking.com. Особенностью отобранных текстов является то, что представленная в них информация носит, преимущественно, субъективный характер, т.е. содержит оценочную и эмоциональную составляющие. Основываясь на личном опыте, пользователи отмечают как положительные, так и отрицательные стороны заданного объекта (*hotel* или *restaurant*), выражая к нему свое личное отношение.

В ходе анализа отобранных текстов для объекта *restaurant* были выделены такие аспекты (характеристики объекта) как *common*, *cuisine*, *interior*, *service*, *price*, а также аспектные термины (характеристики аспекта), например, *atmosphere*, *conditions*, *food*, *bar*, *Wi-fi* и др. Определенные аспекты и аспектные термины были выявлены и для объекта *hotel*. Включение аспектов и аспектных терминов в лингвистическую базу данных помогает исключить те лексические единицы, которые не влияют на оценку объекта. Кроме того, был составлен словарь оценочной лексики, являющийся словарем словоформ и состоящий из двух частей: перечня положительных оценочных слов и перечня отрицательных оценочных слов. Каждая лексическая единица наделена определенным семантическим весом от +3 до –3. Выделение и классификация оценочной лексики проводились с опорой на словари и знание английского языка. Исследование массива текстов отзывов англоязычных пользователей о предприятиях гостиничного и ресторанного бизнеса Беларуси показало, что их авторы применяют различные интенсификаторы, т.е. лексические единицы, увеличивающие или уменьшающие вес оценочного слова, например, *very, highly, extremely, too, badly*. Выявленные интенсификаторы были ранжированы по следующей шкале: высокая, средняя и низкая степень усиления/уменьшения веса оценочного слова. Были также определены слова-инверторы, меняющие направление оценочного веса слова на противоположное. К ним относятся местоимения, предлоги и частицы, например, *delicious (+2)* – ***less*** *delicious (-2), special (+1)* – ***nothing*** *special (-1), bad (–1)* – ***not*** *bad (+1), regret (–1)* – ***without*** *regret (+1)*. К графическим способам усиления/уменьшения оценки были отнесены прописные (заглавные) буквы, восклицательный знак, смайл в виде открывающей или закрывающей круглой скобки (имитирующий улыбку либо огорчение автора).

Выявленные в полном объеме языковые и невербальные маркеры сформировали лингвистическое обеспечение системы автоматического анализа общественного мнения об организации.

Е. Сосновская

СРАВНИТЕЛЬНАЯ ОЦЕНКА РЕЗУЛЬТАТОВ АВТОМАТИЧЕСКОГО АНГЛО-РУССКОГО ПЕРЕВОДА НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

Машинный перевод является одним из самых перспективных и прогрессирующих направлений в области прикладной лингвистики. В мировой практике создания систем автоматического перевода (АП) уже осознана необходимость включения в них экстралингвистических знаний, что вплотную сближает проблемы машинного перевода с проблематикой систем искусственного интеллекта. Многие задачи автоматической обработки текста, в том числе АП, могут частично решаться с использованием статистических данных, но очень важно совмещать статистический анализ с контекстным и чисто лингвистическим. Например, при выборе слова из ряда синонимов, система может обратиться к частотному словарю, чтобы определить, какое сочетание слов встречается чаще в данной тематике.

Для выявление степени эффективности и качества отдельных систем машинного перевода был проведен эксперимент по переводу англоязычного научно-технического текста (текст инструкции) на русский язык. Для тестирования были выбраны наиболее популярные системы машинного перевода Promt и Google Translate.

Как известно, англо-русское и русско-английское направления разработаны довольно тщательно, но несмотря на это возникают проблемы с правильностью перевода. Полученные обеими системами переводы нельзя считать удовлетворительными, так как в каждом из текстов допущены отдельные ошибки, которые искажают смысл высказываний. Системы не совсем верно перевели слова, значение которых можно установить из научной тематики текста*: splitting* ‘разделение’, *Nuclear fission* ‘ядерное разваливание’. Во многих предложениях обе системы неверно осуществили синтаксический анализ и синтез, не согласовали некоторые слова, что противоречит всем правилам синтаксиса: *As a result of fission, other reaction products may also arise* ‘В результате расщепления могут также возникнуть другие продукты реакция’.

Таким образом проанализировав переводы, мы убедились, что в англо-русском направлении научно-технического перевода основными проблемами, влияющими на смысл полученного перевода, является выбор правильного перевода многозначных слов и правильный синтаксический синтез.

В заключении следует отметить, что совершенствование систем машинного перевода происходит довольно активно, внедряются дополнительные блоки и модули, задается ориентация на прагматический и экстралингвисти-