А. С. Скворцова (Минск, МГЛУ)

ФОРМАЛЬНАЯ МОДЕЛЬ СИСТЕМЫ АВТОМАТИЧЕСКОГО АНАЛИЗА ОБЩЕСТВЕННОГО МНЕНИЯ ОБ ОРГАНИЗАЦИИ

В статье рассматриваются принципы организации формальной модели автоматического анализатора общественного мнения о предприятиях гостиничного и ресторанного бизнеса Республики Беларусь. Репутация предприятия определяется компьютером на основе лингвистической базы данных, включающей списки аспектных категорий и аспектных терминов двух анализируемых объектов, а также списки ключевых слов, оценочной лексики, слов-интенсификаторов, слов-инверторов и графических эмотивных символов. Разработанная формальная модель проводит поэтапную обработку текста отзыва о конкретном предприятии на основе базы данных и совокупности определенных лексических и синтаксических правил. Компьютерный эксперимент показал, что положенный в основу формальной модели лексический подход, в целом, позволяет правильно извлекать оценочные суждения пользователей о предприятии и на их основе формулировать вывод о его репутации. В то же время допущенные анализатором ошибки позволяют сформулировать пути совершенствования представленной формальной модели.

Под формальной моделью какого-либо объекта (явления) понимается некоторая система правил, имитирующая его структуру и/или поведение и позволяющая хотя бы частично воспроизвести его либо с помощью человека, либо с помощью компьютера. Для воспроизведения лингвистического объекта (явления) с помощью компьютера необходимо составить базу формализованных данных/знаний, описывающих этот объект (явление), и построить на ее основе алгоритм его функционирования. В статье описывается формальная модель системы автоматического анализа общественного мнения об отеле или ресто-

ране на основе изучения специфики выражения мнений (оценок) в 500 текстах отзывов англоязычных пользователей интернета, размещенных на специализированных сайтах отзывов *http://www.tripadvisor.com* и *http://www. booking.com.*

В ходе анализа отобранного массива текстов для объекта *hotel* и объекта *restaurant* были выделены наборы аспектов (аспектных категорий), а также их атрибуты (аспектные термины). Так, для объекта *hotel* были определены аспектные категории *common* (аспектные термины *atmosphere, location, noise*), *service* (аспектные термины *quality, speed of service, staff*), *room* (аспектные термины *size, design, comfort, noise*), *cuisine* (аспектные термины *quality, taste, size of portions*) и *price* (аспектные термины *level of prices, quality-price relation*). Для объекта *restaurant* были определены следующие аспектные категории: *common* (аспектные термины *atmosphere, location, noise*), *cuisine* (аспектные термины *quality, taste, size of portions*), *service* (аспектные термины *quality, speed of service, staff*) и *price* (аспектные термины *level of prices, quality-price relation*). Для каждой аспектной категории были выделены ключевые слова, с которыми связаны оценочные единицы. Например, к аспекту *cuisine* относятся ключевые единицы *food, meal/meals, dish/dishes, breakfast, lunch, dinner, supper.* Далее на основе текстов отзывов был составлен список оценочной лексики (эмоционально-оценочных и эмоционально-экспрессивных слов). Он представляет собой перечень словоформ и состоит из двух частей: в первой части содержатся слова с положительной оценочной семантикой, во второй – лексические единицы с отрицательной оценочной семантикой. Каждая словоформа наделена определенным семантическим весом от +3 (сверхположительное оценочное значение) до –3 (сверхотрицательное оценочное значение). Например, *The service is* ***excellent*** *(+3). We had a* ***nice*** *(+1) dinner at this restaurant.* Исследование отобранного массива текстов отзывов показало, что их авторы применяют различные интенсификаторы, т.е. лексические единицы, увеличивающие или уменьшающие вес оценочного слова, например, *very, highly, extremely, too, badly.*

Так, положительный интенсификатор, стоящий перед положительным оценочным словом, увеличивает вес данного слова на 1. Например: *We had a* ***really*** *wonderful (+3) dinner.* Этот же интенсификатор, стоящий перед отрицательным оценочным словом, уменьшает вес данного слова на –1: *This restaurant is* ***total*** *disappointment (–4).* Отрицательный интенсификатор влияет на вес оценочного слова в обратном порядке. В результате анализа материала исследования также были выявлены слова-инверторы. К ним относятся местоимения, предлоги и частицы, меняющие направление оценочного веса слова на противоположное. Например, *delicious (+2) –* ***less*** *delicious (–2), special (+1) –* ***nothing*** *special (–1), bad (–1) –* ***not bad*** *(+1), regret (–1) –* ***without*** *regret (+1).* Кроме того, в раз-мещенных на сайтах текстах отзывов встречаются эмотивные графические символы: прописные (заглавные) буквы, восклицательный знак, левая круглая скобка (условно обозначающая улыбку – положительная оценка), правая круглая скобка (условно обозначающая неодобрение автора – отрицательная оценка), изменяющие семантический вес оценочного слова так же, как слова-интенсификаторы. Например, *Breakfast was* ***EXCELLENT*** *(+4). Awesome (+4) hotel*! В качестве инверторов были также рассмотрены двойные и одинарные кавычки. Так, если слово с положительным семантическим весом заключено в кавычки, то направление его семантического веса меняется на противопо-ложное. Например, *The waiters were "friendly" (–1). This hotel was a "pleasure" (–1) to stay in.* Рассмотренные выше типы языковых

и графических маркеров выражения оценки (мнения) позволили сформировать списки единиц, составившие лингвистическую базу данных, с опорой на которую была разработана формальная модель (алгоритм) автоматического анализатора. Рассмотрим основные особенности данного алгоритма.

Формальная модель автоматического анализатора общественного мнения об организации строится на поэтапной обработке текста. На первом этапе из массива текстов (*M*) выбирается очередной текст отзыва о конкретном предприятии. Он разбивается на предложения, каждое из которых анализируется отдельно по словам. Каждое выделенное слово сравнивается с единицами лингвистической базы данных. Если выделенное слово является аспектной категорией, аспектным термином либо ключевым словом, то его ближайший контекст (одно-два слова) проверяется на наличие в списке оценочной лексики. Если аспектная категория либо ключевое слово в предложении найдено не было, то его дальнейший анализ не проводится. Для достижения максимального результата при извлечении оценочных слов анализируется синтаксическая структура предложения. При этом учитываются характеристики, выраженные посредством вспомогательного глагола: аспектная категория/ключевое слово + *is/are/was/were/have/has/will/shall/been* + характеристика. Согласно списку оценочной лексики, в счетчики добавляются веса для каждого аспекта и связанных с ним ключевых слов. Сочетание *аспектная категория / аспектный термин / ключевое слово + оценочная единица* заносится во временный файл.

На втором этапе происходит проверка наличия перед оценочными единицами слов-интенсификаторов. Каждая ветвь проверки наличия интенсификаторов заканчивается проверкой наличия перед ними слов-инверторов. Если перед оценочным словом есть инвертор, но нет интенсификатора, то в соответствии со шкалой оценки его вес меняется на противоположный. Если же перед оценочным словом есть как интенсификатор, так и инвертор, то меняется как вес этого слова, так и при необходимости направление оценки. Если оценочное слова заключено в одинарные или двойные кавычки, то направление его семантической оценки меняется на противоположное. Аналогично проводится проверка на наличие в предложении текста отзыва упомянутых выше графических способов усиления оценки. Если оценочное слово написано прописными (заглавными) буквами, или в конце предложения стоит восклицательный знак, то семантический вес оценочного слова увеличивается на +1 (для лексики с положительной оценочной семантикой) либо на –1 (для лексики с отрицательной оценочной семантикой). Таким же образом на семантический вес слова влияет наличие открывающей круглой скобки в предложении (семантический вес слова увеличивается на –1) либо закрывающей круглой скобки (семантический вес слова увеличивается на +1).

На третьем этапе осуществляется выделение двухкомпонентных лексических единиц. База данных системы содержит лексические единицы, которые являются оценочными либо интенсификаторами только при сочетании двух элементов, например, *all* ***right***, ***thumbs*** *up*, *at* ***all***, *by* ***far*** (выделен главный элемент словосочетания). Для таких единиц вводятся следующие правила. Если проверяемое слово – *right*, а его левый контекст – *all*, то единица является оценочной. Если проверяемое слово *thumbs*, а его правый контекст *up*, то единица является оценочной. Если проверяемое слово *all*, а его левый контекст *at*, то единица является интенсификатором. Если проверяемое слово *far*, а его левый контекст *by*, то единица является интенсификатором.

На четвертом этапе компьютер суммирует показатели счетчиков для каждой аспектной категории в общий счетчик веса ее положительной или отрицательной оценки для конкретного текста отзыва о гостинице или ресторане. Сравнивая общие счетчики положительной и отрицательной оценки, система делает вывод о направлении оценки данного аспекта. Вместе с извлеченной из временного файла информацией окончательный результат анализа отдельного текста отзыва о предприятии имеет следующий вид:

Object: Name of Hotel
Review about hotel (text)
Total sentences – (number)
Evaluation (total) – (positive number / negative number)
Summary *– (positive / neutral / negative)*
Sentence details
Sentence 1
Sentence with keyword(s)
Total keywords: – (number)
Aspect – (aspect category / aspect term)
Keyword – (keyword name)
Evaluator – (evaluator name)
Intensifier – (intensifier name)
Inverter – (yes/no)
Evaluation *– (positive number / negative number).*

На пятом этапе после анализа всех текстов отзывов о конкретном предприятии компьютер суммирует показатели всех счетчиков по аспектным категориям в единый счетчик веса положительной оценки и единый счетчик веса отрицательной оценки. Сравнивая счетчики положительных и отрицательных весов, система делает окончательный вывод об общественном мнении о конкретном предприятии. Так, итоговый результат анализа общественного мнения о гостинице имеет следующий вид:

Aspect "Common" – (positive / neutral / negative)
Aspect "Service " – (positive / neutral / negative)
Aspect "Room" – (positive / neutral / negative)
Aspect "Cuisine" – (positive / neutral / negative)
Aspect "Price" – (positive / neutral / negative)
Number of review texts (items) – (number)
Resume: public opinion *– (positive / neutral / negative).*

На основе лингвистической базы данных и формальной модели системы автоматического определения репутации предприятия был создан компьютерный анализатор (язык программирования Python). В ходе компьютерного эксперимента анализатор обработал несколько сотен текстов отзывов о предприятиях указанного типа. Проведение компьютерного эксперимента и анализ его результатов позволили сделать следующие выводы и определить пути совершенствования разработанной формальной модели.

1. Компьютер может определить наличие в тексте оценочных слов, их вес и сделать вывод об общем направлении оценки (положительная или отрицательная) для каждой присутствующей в тексте отзыва аспектной категории объекта *hotel* и объекта *restaurant*.

2. В процессе извлечения оценочных суждений из текста отзыва наиболее сложной проблемой является правильное определение синтаксических структур предложений. Это связано с тем, что анализируемый материал представляет собой тексты, язык которых можно определить как разговорный письменный. Авторы таких текстов не всегда соблюдают расстановку знаков препинания, либо игнорируют их вообще, а также допускают орфографические ошибки. Наличие неправильно написанных слов мешает правильному анализу текста, выделению, как оценочных слов, так и объектов оценки. Поэтому в разработанную формальную модель необходимо включить модуль нормализации лексических единиц, который позволит частично решить указанную проблему.

3. На точность процедуры правильного выделения оценочных слов значительное влияние оказывает омонимичность форм слов и частей речи, например, *like* – глагол (оценочный) и *like* – союз (не оценочный), *little* – прилагательное (оценочное) и *little* – наречие (интенсификатор). Для решения данной проблемы нужно проводить дополнительную процедуру тегирования текста по частям речи.

4. Отдельно стоит отметить употребление некоторых оценочных слов. Например, положительное слово *high* (+1) в сочетании с названием аспектной категории *price* меняет направление оценки на противоположное. Необходимо выявить все подобные случаи и разработать дополнительные правила лексической сочетаемости, влияющие на определение веса оценки.

5. Компьютер не может правильно извлекать информацию из сложных синтаксических конструкций, основываясь только на правилах поиска оценочных слов в ближайшем контексте ключевого слова конкретной аспектной категории. Например, при обработке предложения *The breakfast had everything that could be wished* система правильно распознает аспектную категорию *cuisine* по ключевому слову *breakfast*, однако не может правильно определить ее оценочность, так как в данном случае она выражена не лексической единицей, а синтаксической конструкцией *had everything that could be wished*. Поэтому предложенная формальная модель может быть дополнена рядом синтаксических правил, учитывающих синтаксическую структуру оценочных конструкций. Таким образом, при пополнении лингвистической базы данных и расширении формальной модели путем добавления дополнительных лексических и синтаксических правил анализатор будет точнее выделять аспектные категории, аспектные термины и оценочные слова, а значит, во всех случаях корректно определять репутацию организации.

Разработанная база данных, формальная модель и сделанные по итогам компьютерного эксперимента выводы могут стать основой для создания нейронной сети, которая будет самообучаться и гораздо эффективнее распознавать оценочные единицы с целью последующего формулирования на их основе общественного мнения.

The article deals with the main principles of organizing a formal model of an automatic analyzer of public opinion about hotels and restaurants of the Republic of Belarus. The developed formal model carries out step-by-step processing of a review text taking into account a database and a set of lexical and syntactic rules. The computer experiment showed that the lexical approach made it possible to extract users' opinions about a hotel or a restaurant correctly and to formulate a conclusion about its reputation. At the same time, the errors made by the analyzer allowed to formulate some ways to improve the formal model.