А. В. Молоткова
Минск, МГЛУ

ФОРМАЛЬНАЯ МОДЕЛЬ ПРОЦЕССА АКТУАЛИЗАЦИИ ПРЕДМЕТНОЙ БАЗЫ ДАННЫХ О ДТП (лингвистический аспект)

В статье рассматриваются основные принципы лингвистической организации формальной модели процесса актуализации предметной базы данных о дорожно-транспортных происшествиях на территории Испании. В частности отмечается, что процесс актуализации предметной базы данных включает ввод поискового запроса пользователя, извлечение из предметной базы данных информации, релевантной запросу пользователя, и печать результатов поиска. Описывается структура и процедура обработки следующих типов запросов пользователя к предметной базе данных: простого запроса, сложного запроса общего типа и сложного запроса конкретного типа. Все запросы формулируются на естественном языке.

Несмотря на разнообразие существующих методов извлечения информации из различных типов массивов данных, до сих пор не решены ключевые проблемы информационного поиска, связанные с автоматическим построением предметных баз данных и знаний. Автором статьи была сделана попытка смоделировать процесс автоматического извлечения именованных сущностей и фактов из текстов испаноязычных новостных сводок о дорожно-транспортных происшествиях (ДТП) с целью формирования и последующей актуализации предметной базы данных (ПБД). Именованной сущностью считается слово или словосочетание, предназначенное для конкретного, вполне определенного предмета или явления, выделяющее этот предмет или явление из ряда однотипных предметов или явлений. Именованная сущность обязательно имеет референт, то есть объект внеязыковой действительности, подразумеваемый автором конкретного речевого отрезка. В письменном тексте именованные сущности представлены объектами и их атрибутами, выраженными определенными лексическими и фразовыми шаблонами. В зависимости от предметной области и тематики текста в нем используются факты и их атрибуты, представленные определенным пластом лексики.

Материалом исследования послужили 50 текстов новостных сводок о ДТП, взятых с различных испаноязычных сайтов Интернета. В ходе анализа отобранных текстов в качестве объектов рассматривалась, во-первых, персона, характеризующаяся такими атрибутами, как количество участвовавших в ДТП персон, возраст персоны, состояние персоны после аварии; во-вторых, локация, указывающая на конкретное место происшествия и название дороги; в-третьих, организация, имеющая такие атрибуты, как название организации и адрес; в-четвертых, марка автомобиля. В новостных сводках о дорожно-транспортных происшествиях были выявлены следующие факты: дата и время инцидента, тип транспортного средства, тип проис-шествия, причина аварии и результат аварии. В процессе выполнения исследования были выделены определенные лексические и фразовые шаблоны, послужившие основой для извлечения объектов, фактов и их атри-

бутов из испаноязычных текстов о ДТП. Также были составлены списки лексических единиц, обозначающих такие факты, как тип транспортного средства и причина аварии, в связи с тем, что конкретные маркеры, указывающие на данные факты, в текстах сводок о ДТП отсутствуют. Лексические и фразовые шаблоны, списки конкретных лексических единиц сформировали лингвистическое обеспечение системы автоматической обработки текстов подобного типа.

На основе разработанной лингвистической базы данных была построена формальная модель, состоящая из двух частей. Первая часть отражает особенности процесса извлечения объектов, фактов и их атрибутов из текстов испаноязычных новостных сводок о дорожно-транспортных происшествиях с целью формирования предметной базы данных. В основу работы второй части формальной модели, связанной с автоматической актуализацией созданной ранее предметной базы данных о ДТП на территории Испании, положен набор формализованных правил. Рассмотрим их подробнее.

Процесс актуализации предметной базы данных состоит из следующих этапов: 1) ввода поискового запроса пользователя; 2) извлечения из предметной базы данных информации, релевантной запросу пользователя и 3) печати результатов поиска. При обращении к предметной базе данных можно использовать несколько типов поисковых запросов. Первый тип представляет собой простой запрос, то есть поисковой запрос в виде одного параметра, указывающего на одно поле ПБД. Например, в качестве запроса пользователь может ввести только дату ДТП либо только место дорожно-транспортного происшествия. В результате обработки такого запроса ему будут выданы все записи предметной базы данных, удовлетворяющие этому запросу. Второй тип представляет собой сложный запрос, основанный на нескольких полях ПБД. Сложный запрос может быть общим и конкретным. Сложный запрос общего типа представлен в виде нескольких параметров поиска, то есть может одновременное указывать на следующие поля предметной базы данных: дата/день недели ДТП, место ДТП, название дороги, на которой произошло ДТП, тип ДТП. Например, поиск может осуществляться одновременно по заданной дате и месту происшествия, либо по дате и названию дороги, на которой произошло дорожно-транспортное происшествие, либо по дате, месту и названию дороги. После обработки такого запроса пользователю будут выданы все записи предметной базы данных, удовлетворяющие этому запросу, однако результат поиска по сложному запросу общего типа будет значительно уже результата поиска по простому запросу. Сложный запрос конкретного типа также представлен в виде нескольких параметров поиска, однако предполагает одновременное указание на несколько полей ПБД, среди которых обязательно должно быть поле, отражающее детальную информацию о дорожно-транспортном происшествии. После обработки такого запроса пользователю будут выданы не все записи предметной базы данных, удовлетворяющие этому запросу, а значение конкретного данного, то есть результат поиска по сложному запросу конкретного типа будет значительно уже результата поиска по сложному

запросу общего типа. Например, пользователь может сформулировать конкретный запрос о количество участников ДТП, которые произошли в конкретный день и в конкретном месте, их состоянии после аварии, месте, куда участники происшествия были доставлены и т.п. Таким образом, в качестве основных параметров, как минимум один из которых должен присутствовать в поисковом запросе, отметим следующие: дата/день недели, место дорожно-транспортного происшествия, название дороги, тип ДТП. К дополнительным отнесем параметры, непосредственно связанные с участниками дорожно-транспортного происшествия – их количеством, возрастом, состоянием после аварии, названием и адресом организации, куда были доставлены пострадавшие, а также информацию о времени ДТП, типе транспортного средства, марке транспортного средства, организации, оказавшей пострадавшим помощь и т.п.

Рассмотрим основные особенности второй части формальной модели на конкретных примерах. Как отмечалось выше, в любом новостном сообщении о дорожно-транспортном происшествии указана его дата. Следовательно, задав поиск по дате, то есть указав в качестве простого поискового запроса число, месяц и год, например, *16/04/2016, 14 de marzo de 2015*, пользователь получит все имеющиеся в предметной базе данных сведения об авариях, произошедших в этот день на дорогах Испании. Если пользователь вводит название области, провинции или города, который его интересует с точки зрения произошедших в нем ДТП, например, *Andalucía*, *Madrid*, *Burgos*, то в результате актуализации ПБД пользователь получит все имеющиеся данные о дорожно-транспортных происшествиях, которые произошли в указанном месте. Таким же образом можно получать информацию о всех ДТП, которые произошли на определенной дороге Испании, указав в качестве запроса на поиск ее название, например, *la carretera N-I*, *el kilómetro 275 de la N-I*, а также сортировать происшествия по типам, например, *salidas de la vía*, *vuelco y pérdida de control*, *atropellos*, *colisiones entre dos vehículos*, *colisiones múltiples o en cadena*.

Для извлечения из предметной базы данных более детальной информации следует формулировать сложный поисковый запрос общего типа, то есть запрос, основанный на двух и более основных параметрах. Так, указав в качестве параметров поиска дату *16/05/2017* и место происшествия *Burgos*, пользователь получит всю информацию о ДТП, которые произошли в данный день на данной территории. Для того чтобы сузить результаты поиска, к уже заданным параметрам *16/05/2017, Burgos* необходимо добавить другие параметры, например, указать название дороги *carretera N-1*. Таким образом, пользователь получит информацию о дорожно-транспортных происшествиях, которые произошли 16 мая 2017 года в г. Бургосе на трассе Н-1. При обработке простого запроса и сложного запроса общего типа в результате актуализации предметной базы данных компьютерная система предоставляет пользователю все записи, которые входят в ПБД и соответствуют введенному запросу.

Если пользователю требуется получить конкретную информацию о дорожно-транспортном происшествии, формулируется сложный поисковый запрос конкретного типа. В этом случае пользователь указывает, какие конкретные объекты, факты или их атрибуты необходимо извлечь из предметной базы данных, например, количество участников ДТП, их состояние, тип транспортного средства и т.п. При обработке сложного запроса конкретного типа в результате актуализации предметной базы данных компьютерная система предоставляет пользователю не все записи, которые входят в ПБД, а отображает только те данные, которые запросил пользователь.

Взаимодействие пользователя с предметной базой данных осуществляется посредством дисплея, на который выводятся названия всех полей ПБД. Таким образом, компьютер предлагает пользователю выбрать поле предметной базы данных, по которому будет производиться поиск интересующей его информации. После нажатия на какое-либо из предложенных в интерфейсе полей появляется дополнительное окно ввода, в которое пользователь может ввести соответствующий поисковый запрос (рис. 1).

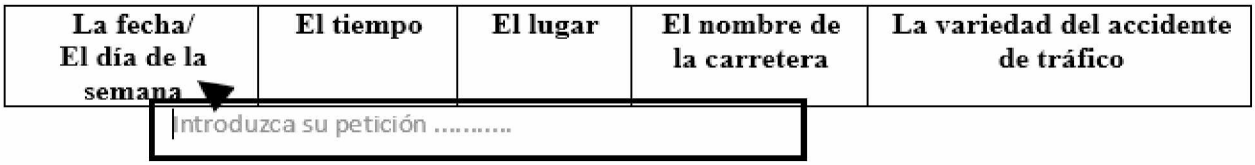

Рис. 1. Фрагмент интерфейса системы на этапе актуализации ПБД с помощью простого запроса

Пользователь вводит необходимый ему запрос/запросы. Далее система уточняет, интересует ли пользователя конкретная информация (рис 2). Если пользователь отвечает положительно, компьютерная система предлагает выбрать поле ПБД, по которому необходимо предоставить информацию. В таком случае, после завершения поиска, система выдаст на печать только те данные, которые соответствуют сделанному запросу. Если его интересует вся информация, то система выдает на печать все данные, связанные с дорожно-транспортными происшествиями, которые соответствуют введенным запросам. Таким образом, пользователь сам определяет параметры поиска информации в ПБД.

Проверка адекватности формальной модели и анализ ее результатов позволяет сделать следующие выводы. В целом компьютерная система способна правильно определять и выделять объекты, факты и их атрибуты из текстов испаноязычных новостных сводок о дорожно-транспортных происшествиях. Однако правильное выделение ряда данных может вызвать у компьютерной системы затруднения. Так, сложность может представлять определение количества участников ДТП. Это связано с особенностями грамматических структур (*uno de los dos*), а также с наличием омонимичных форм слова и частей речи, например, *un* – неопределенный артикль и *un* – усеченная форма числительного *uno* перед существительными. Решение данной проблемы видится в автоматическом тегировании текстов по частям речи.

Кроме того, на точное определение количества участников аварии влияет синонимичность слов. Если система не сможет определять синонимы, то все лексические единицы, соответствующие объекту *persona*, будут обозначать участников происшествия. Решение этой проблемы видится в разработке.

La fecha / El día de la semana	El tiempo	El lugar	El nombre de la carretera	La variedad del accidente de tráfico
13/08/2016		*Sevilla*		

¿Es necesario buscar información específica?

Sí No

La fecha/El día de la semana
La hora
El lugar
El nombre de la carretera
La variedad del accidente de tráfico
La variedad de vehículo
La marca del vehículo
La cantidad de los participantes
La persona
La edad
El estado
La organización de urgencia
La ubicación

Рис. 2. Фрагмент интерфейса системы на этапе актуализации ПБД с помощью сложного (конкретного) запроса

Большую роль в точном извлечении объектов, фактов и их атрибутов играют синтаксические структуры предложений новостных сообщений. Например, в предложениях часто опускается подлежащее, либо оно заменяется указательным местоимением (*este*, *aquel*), что в свою очередь может привести к неточности при извлечении нужной информации. Для улучшения работы формальной модели формирования и актуализации предметной базы данных необходимо расширить списки лексических единиц и перечни шаблонов, указывающих на объекты, факты и их атрибуты. Это приведет к более точному извлечению данных из новостных сводок о дорожно-транспортных происшествиях и, соответственно, к более качественному формированию предметной базы данных.

The article deals with specific linguistic features of the formal model of automatic actualization of road accidents object database. Different types of user's queries to the database are under consideration.

В. Н. Мурашко
Минск, МГЛУ

ИМЕНОВАННАЯ СУЩНОСТЬ И ЕЕ КАТЕГОРИИ

В статье рассматривается понятие именованной сущности и основные категории, к которым ее можно отнести: человек, место, организация, дата и время. Отмечается, что выбор соответствующей категории не всегда определяется самой сутью именованной сущности. На примерах показано, что в ряде случаев категория зависит от микроконтекста ее употребления. В работе указаны основные трудности, с которыми может столкнуться система автоматического распознавания именованных сущностей в письменном тексте.

Под именованной сущностью понимают «что-либо реально существующее или вымышленное, на что можно указать или к чему можно обратиться при помощи имени собственного» [1, с. 214]. В соответствии с этим определением в задачу распознавания именованных сущностей входит не только нахождение их в тексте, но и однозначное указание на подразумеваемый объект или лицо, а также приписывание ему определенной категории. Чаще всего используется простая классификация, включающая в себя следующие три категории: ЧЕЛОВЕК (сокращенно ПЕР от *персона*), МЕСТО (сокращенно ЛОК от *локация*), ОРГАНИЗАЦИЯ (сокращенно ОРГ) [2, p. 1538]. Такая классификация является довольно простой, поскольку имеются задачи, где важно различать подтипы именованных сущностей. Например, если идет речь о компании или стране, об актере или политике и т.п. Однако разработать более подробную схему, подходящую для текстов разных жанров, намного сложнее. Хотя ни даты, ни числа не соответствуют приведенному выше определению, они часто распознаются как именованные сущности вместе с ПЕР, ЛОК и ОРГ. Для обозначения этих категорий используются сокращения ТЕМП (темпоральные выражения) и НУМ (нумерические выражения) [Там же, p. 1539]. Следует отметить, что, в отличие от трех других категорий, распознавание чисел и времени представляется значительно более легкой задачей, поскольку существует ограниченный набор способов их выражения. Ниже в таблице приведены различные именованные сущности и соответствующие им категории, содержащиеся в примере, взятом из Википедии: *Современный* [***СПбГУ***] *в* [***России***] *– преемник* [***Академического университета***], *который был учрежден одновременно с* [***Академией наук***] *указом* [***Петра I***] *от* [***28 января 1724 года***]. *В частности, в* [***1758–1765***] *годах ректором* [***Академического университета***] *был* [***М. В Ломоносов***].

Как видно из таблицы, помимо очевидной категории «человек», М. В. Ломоносову можно приписать и другие категории, соответствующие его многогранной личности. В данном случае речь идет не о более подробной