тителей «посадочной» страницы является отзыв благодарных покупателей. Отзывы должны быть конкретными. Конкретные числа, реальные данные и описание определенного применения предмета рекламы нравятся пользователям. Хорошим способом заставить пользователя быстрее совершить нужное рекламодателю действие является использование популярных психологических триггеров – срочности (ограничение предложения по времени) и дефицита (ограничение количества товара). Чтобы привлечь внимание пользователей к самым важным элементам страницы, следует воспользоваться стрелками. Важным визуальным компонентом целевой страницы является свободное пространство. В этом случае она будет выглядеть сбалансированной, а кнопка с призывом к действию выделяться среди других элементов. Общеизвестно, что лучший способ привлечь внимание и подчеркнуть ценность предмета рекламы заключается в использовании изображения (мозг обрабатывает изображения в 60 000 раз быстрее, чем текст). Фотографии должны иметь прямое отношение к товару или услуге.

Таким образом, единого шаблона идеальной «посадочной» страницы не существует. Необходимо каждый раз адаптировать дизайн, изображения и текст под конкретную целевую аудиторию и тематику, а также под канал, с которого идет трафик. Все это способствует улучшению организации целевой страницы и увеличению процента реальных действий или продаж, т.е. конвертации посетителей страницы в потенциальных покупателей.

ЛИТЕРАТУРА

1. *Петроченков, А. Е.* Идеальный Landing Page. Создаем продающие веб-страницы / А. Е. Петроченков. – М. : Экспо, 2017. – 409 с.
2. *Толмачев, А. А.* Реклама в Интернет. Курс молодого бойца / А. А. Толмачев. – М. : BHV, 2015. – 240 с.
3. *Голополосов, Д. В.* Способы повышения конверсии сайтов / Д. В. Голополосов. – М. : Издательский дом «Питер», 2013. – 160 с.

The article deals with a new kind of network advertising for the Belarusian market – a landing page. It's noted that this type of advertising allows to segment the Internet users correctly, effectively influencing them and converting quickly in potential consumers of goods or services.

В. Н. Мурашко (*Минск, МГЛУ*)

МОДЕЛИРОВАНИЕ ПРОЦЕССА АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ ФРАНКОЯЗЫЧНЫХ НОВОСТНЫХ СООБЩЕНИЙ

В статье рассматривается структура формальной модели автоматического извлечения именованных сущностей из текстов франкоязычных СМИ. Отмечается, что в письменном тексте именованная сущность выражается именем собственным, указывающим на конкретного человека, организацию, географический объект и т.д. и относящимся к опре-

деленной категории. В процессе анализа материала исследования были выявлены именованные сущности таких категорий, как персона, организация, географический объект, геополитический объект, дата, произведение искусства, а также их всевозможные аспекты или атрибуты. Кроме того, были выделены лексические и грамматические маркеры, фразовые шаблоны, а также правила определения левой и правой границы именованных сущностей в тексте. Конкретные примеры иллюстрируют принципы организации лингвистической базы данных как первой части формальной модели. Приводится описание алгоритма автоматического выполнения указанной процедуры (второй части формальной модели). Его работа демонстрируется на примере обработки текста франкоязычного новостного сообщения компьютером.

Новостные сообщения, освещающие относящиеся к политической, социальной, экономической и другим сферам события, являются хорошим источником, из которого автоматически можно извлекать несколько типов данных. Так, в зависимости от предметной области и тематики в тексте новостного сообщения используются связанные с представленными событиями определенные именованные сущности, то есть имена и фамилии людей, названия организаций, географических и геополитических объектов и т.д. В целом под *именованной сущностью* понимают реально существующий или вымышленный объект, на который можно указать или к которому можно обратиться при помощи имени собственного. Таким образом, именованная сущность обязательно имеет референт, подразумеваемый автором текста. Референт может принадлежать реальному миру, например, фамилия, имя и отчество конкретного человека или название конкретной организации, либо вымышленному миру, например, это может быть персонаж художественного произведения. Поэтому автоматическое распознавание именованной сущности в тексте новостного сообщения связано не только с ее выявлением, но и с приписыванием ей определенной категории, то есть с определением однозначного указания на подразумеваемый объект или лицо.

Рассмотрим моделирование процесса автоматического извлечения именованных сущностей из франкоязычных новостных сообщений на основе инженерного подхода, опирающегося на заранее созданные словари и определенные правила. Под формальной моделью какого-либо лингвистического 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 ходе анализа 150 текстов новостей, взятых с сайта *fr.euronews.com*, в качестве искомых объектов были определены именованные сущности таких категорий, как *персона, организация, географический объект, геополитический объект, дата, произведение искусства*, а также их всевозможные аспекты или атрибуты (*должность, звание, профессия; партия, фирма, корпорация, медиаорганизация, банк, комиссия; континент, страна, столица, город, район, океан, озеро, море, река, остров, гора, горная цепь, пустыня, лес; правительство, народ; опера, балет, фильм, книга, спектакль, песня*).

В процессе анализа отмеченных выше категорий именованных сущностей и их атрибутов были выделены конкретные лексические/грамматические маркеры и фразовые шаблоны, а также ряд правил, позволивших определить левую и правую границы именованных сущностей в текстах данного типа. Рассмотрим принципы организации перечисленных данных на примере именованной сущности *организация*. Главным отличием именованной сущности категории *организация* от именованной сущности категории *персона* является наличие перед ней определенного артикля (маркер левой границы именованной сущности). Например: *Le président de* ***la Commission européenne****, Jean-Claude Juncker, a exclu vendredi de se séparer de son bras droit.* ***Le Conseil de sécurité des Nations unies*** *s'est réuni*. Определенный артикль может быть представлен формами *le*, *la*, *les*, *l'*, *au*, *aux*, *du*, *des* и начинаться со строчной или прописной буквы. Маркером правой границы именованной сущности категории *организация* может служить глагол. Если сказуемое в предложении выражено глаголом в форме прошедшего времени, то грамматическими маркерами будут формы вспомогательных глаголов *avoir* и *être* (*a, est, avait, était, n'a, n'est, n'avait, n'était*), а также их формы для местоименных глаголов (*s'est, se sont, s'était*). В некоторых случаях правой границей категории *организация* может служить предлог, если за ним находится слово, входящее в список географических объектов, например, ***le Tribunal suprême*** *à* ***Madrid***.

В исследованном массиве текстов франкоязычных СМИ именованная сущность категории *организация* имеет такие аспекты, как *партия, фирма, корпорация, медиаорганизация, банк, комиссия, организация*. Маркером данных аспектов может являться первое слово в названии организации, например, *la* ***Commission*** *européenne*, ***Banque*** *de France*. Кроме того, перечисленные аспекты задаются маркерами типа *partie, movement, union, entreprise, firme, banque, bank, journal, réseau social* либо шаблонами вида *la partie *, le mouvement *, l'entreprise *, la companie *, banque *, magazine *, journal **. Маркерами аспектов *корпорация, комиссия* и *организация* служат французские варианты названий самих аспектов: *corporation, commission, organisation*. Необходимо отметить, что аспект *организация* приписывается всем словам, совпадающим по своим характеристикам с именованной сущностью категории *организация*, но не имеющим отмеченных выше маркеров или шаблонов, либо предшествующих слов *corporation, commission, organisation*. Кроме того, в текстах новостных сообщений могут отсутствовать маркеры, прямо указывающие на аспект *медиаорганизация*. Это касается названий известных во Франции и во всем мире газет, журналов, радио и телеканалов, социальных сетей и т.д., поэтому их нужно задавать в базе данных в виде списка лексических единиц. В ходе исследования выяснилось, что в роли названия организации могут выступать различные аббревиатуры. Сразу после аббревиатуры в скобках следует ее расшифровка или аббревиатура может быть указана в скобках после названия организации. Например, *UMP (L'Union pour un mouvement populaire), l'Organisation pour l'interdiction des armes chimiques (OIAC)*. Кроме того, в скобках может быть указан какой-либо комментарий или примечание редакции новостей. Идентифика-

тором того, что в скобках находится расшифровка аббревиатуры, а не комментарий, служит отсутствие смыслового глагола или вспомогательного глагола *avoir* и *être*. Таким образом, левой границей именованной сущности категории *организация* является первое, написанное с прописной буквы слово, а правой границей будет закрывающая скобка. Если в тексте новостного сообщения аббревиатура не имеет какого-либо маркера, она будет отнесена к именованной сущности категории *организация*, а ее аспект не будет определен. Также к именованной сущности категории *организация* будет относиться слово, написанное с прописной буквы, слева от которого стоит определенный артикль, или его слитные формы написания с предлогом *á* и *de*, а справа не будет слова с капитализацией.

Аналогичным образом изучались способы выражения именованных сущностей других категорий. Например, для представления именованной сущности *дата* был выявлен следующий ряд шаблонов. Чаще всего дата публикации новостного сообщения извлекается по шаблону число/месяц/год, например, *25/06/2017*. В тексте самого сообщения месяц может быть указан словесно, а год может не указываться вообще. Поэтому для извлечения именованной сущности категории *дата* был составлен список лексических единиц с указанием месяцев года на французском языке: *janvier, février, mars, avril, mai, juin, juillet, août, septembre, octobre, novembre* и *décembre*. Согласно правилу, при извлечении даты все числовые выражения будут сравниваться с шаблоном **/**/**** (число/месяц/год) либо **/** (число/месяц). При нахождении лексической единицы, входящей в список месяцев на французском языке, осуществляется поиск числительного сначала слева, а затем справа от данного слова. В случае его нахождения в результаты обработки будут включены как месяц, так и найденные числительные.

Выделенные из текстов франкоязычных новостей лексические и грамматические маркеры, а также фразовые шаблоны сформировали первую часть формальной модели – лингвистическую базу данных. На основе базы данных была разработана вторая часть формальной модели – алгоритм, отражающий особенности процесса автоматического извлечения именованных сущностей и их аспектов из текстов франкоязычных новостей. Рассмотрим его основные особенности. Процесс извлечения именованных сущностей из текстов франкоязычных СМИ состоит из двух этапов: ввода текста и непосредственного извлечения из него всех именованных сущностей и их аспектов с выводом полученных результатов на соответствующее устройство (экран или принтер). После ввода текст помещается в специальное поле для обработки и разбивается на отдельные предложения. Для корректной работы системы после знака ’ автоматически ставится пробел. Пробелами также отделяются все знаки препинания (кроме знака –).

Поиск и извлечение именованной сущности опирается, главным образом, на слова, написанные с прописной буквы, а также на такие вспомогательные части речи, как предлоги и артикли. После подготовительной работы система начинает поэтапное выделение слов в рамках одного предложения. В зависи-

мости от первой буквы данного слова производится его сравнение с лексическими маркерами, представленными в лингвистической базе данных и начинающимися со строчной буквы, либо сравнение окружения слова, если оно начинается с прописной буквы.

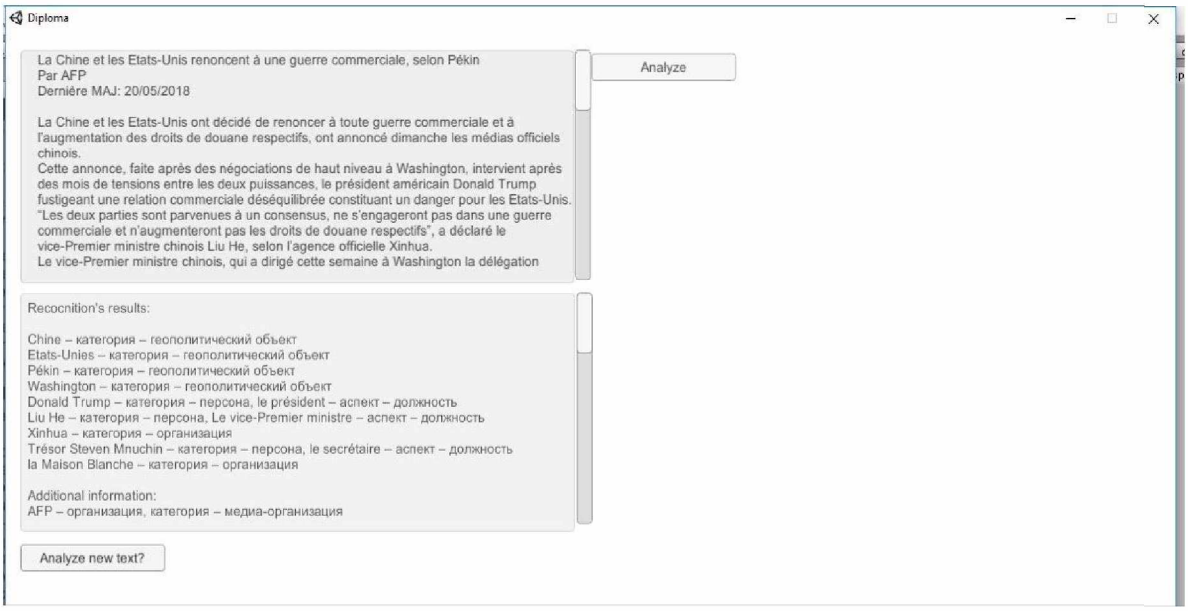
Информация между заголовком и основным текстом новостного сообщения анализируется отдельно. Здесь указывается дата публикации новости (чаще всего, по шаблону **/**/****), а также ее автор, который может быть либо персоной, либо медиаорганизацией. Дата публикации может находиться в позиции после группы слов *Dernière MAJ:*. В таком случае из данного предложения извлекается только дата, а слово *MAJ* игнорируется. Автор новостного сообщения указывается в отдельном предложении, оно всегда начинается с предлога *par*, после которого все слова пишутся с прописной буквы. Для определения точной категории именованной сущности проводится анализ всех слов предложения. Например, если после предлога *par* следует определенный артикль и несколько слов либо одно слово с прописной буквы, то все слова до конца предложения будут отнесены к аспекту *медиаорганизация* категории *организация*. Если после предлога *par* расположено несколько слов без артикля, написанных с прописной буквы, то всем словам справа от предлога до конца строки приписывается категория *персона* и аспект *журналист*. Информация о дате публикации новостного сообщения и его авторе размещается в конце результатов извлечения именованных сущностей с пометкой *Дополнительная информация о новостном сообщении.*

Для проведения компьютерного эксперимента на основе формальной модели был написан программный код на языке программирования ***C#***. Программный продукт функционирует в виде ***.exe*** приложения, созданного при помощи среды ***Unity***. На рисунке представлен фрагмент компьютерного эксперимента.

Результаты компьютерного эксперимента позволяют сделать следующие выводы о возможных путях совершенствования разработанной формальной модели.

1. Для улучшения качества процесса извлечения именованных сущностей необходимо расширить лингвистическую базу данных, а также правила, позволяющие определять именованные сущности категории *организация*, употребленные в тексте без артикля.

2. С целью пополнения лингвистической базы данных лексическими маркерами можно создать дополнительный блок, производящий автоматический поиск и извлечение событий, связанных с упомянутыми в тексте именованными сущностями. Для этого необходимо сформировать списки определенных глаголов и отглагольных имен существительных, которые могут указывать на совершение какого-либо события. Список представленных в данной версии программы глаголов не подходит для решения этой задачи, так как не все эти глаголы являются смысловыми. Например, в предложении ***Jean-Luc Mélenchon*** *ne se sent jamais aussi bien que dans le rôle de trublion* при извлечении именованной сущности категории *персона* глагол играет роль стоп-слова, но не отвечает требованиям по извлечению событий.



Фрагмент компьютерного эксперимента

3. Если пользователь работает с большим количеством текстов новостных сообщений, с целью сохранения извлеченных именованных сущностей можно автоматически создать специальную базу данных. По такой базе данных возможен автоматический поиск именованных сущностей, а также ее автоматическое пополнение. Допустим, в одном тексте какой-либо аспект именованной сущности категории *организация* конкретизирован, а в другом тексте такой конкретизации нет. В первом случае компьютер извлечет аспект самостоятельно, а во втором случае такого действия не произойдет. Данная проблема может решаться путем сравнения результатов обработки большого количества текстов, где недостающие элементы могут быть автоматически найдены и приписаны именованным сущностям из других текстов.

Результаты моделирования процесса автоматического извлечения именованных сущностей из франкоязычных новостных сообщений могут быть использованы для создания промышленной системы поиска и извлечения именованных сущностей разных категорий с различными аспектами из текстов новостей. Они могут стать основой для разработки классификатора, распознающего данные объекты в неструктурированных текстовых массивах сети Интернет, а также для создания нейронной сети, которая будет самообучаться и эффективно распознавать и извлекать именованные сущности, связанные с ними факты и события не только из текстов новостных сообщений, но и из текстов других жанров и стилей.

The article deals with the formal model structure of automatic named entities extraction from French mass media texts. The model includes a linguistic database, an algorithm of the procedure mentioned and a computer program.