Таблица

Поисковый образ для тематики «Film»

act	acts	acting	acted
actor	actors		
contractor	contractors		
director	directors		
drama			
film	films		
finance	financial	financed	
movie	movies		
play	plays	playing	played
star	stars	starring	starred
take	takes	took	taken

На основании разработанных поисковых образов по каждой тематике был составлен алгоритм автоматической идентификации тематики англоязычных публицистических текстов.

The article analyses the procedure of creating a search image for English publicist texts, which is based on the algorithm of identifying key or prop words.

А. О. Третьяк
Минск, МГЛУ

АВТОМАТИЧЕСКОЕ ОБЪЕДИНЕНИЕ ФРАНКОЯЗЫЧНЫХ НОВОСТНЫХ СООБЩЕНИЙ В ЕДИНЫЙ СЮЖЕТ

Формальная процедура автоматического объединения текстов новостей в единый сюжет базируется на решении двух взаимосвязанных задач: определении тематики текстов новостных сообщений на основе дескрипторного метода и автоматическом построении их информационных портретов с опорой на статистический метод. Считается, что несколько новостных сообщений относятся к определенному сюжету при условии сходства их информационных портретов. Последние формируются на основе главных опорных слов каждого текста и формальных синтаксических групп (субъекта, предиката, объекта, места действия, времени действия), выделенных из ключевых предложений новостных сообщений.

Для решения целого ряда аналитических задач возникает потребность оценить содержание текста, получить его информационный портрет, то есть статистически значимую совокупность информационных характеристик. В большинстве существующих систем такой портрет состоит из статистически значимых слов и выражений, сопровождающих упоминание конкретного объекта. Построение информационных портретов позволяет автоматически группировать тексты по сюжетам. В общем плане под сюжетом понимается последовательное

пространственно-временное изложение событий, имеющих определенную тематическую направленность. Под новостным сюжетом понимается совокупность сведений о некоторых сущностях и явлениях: людях, предметах, отношениях, процессах, явлениях, событиях и т.д. Можно предположить, что группа схожих по содержанию (информационным портретам) и близких по времени новостных сообщений будет соответствовать одному новостному сюжету.

В работе рассматривается проблема автоматического объединения франкоязычных новостных сообщений в единый сюжет. Материалом исследования послужили 33 текста франкоязычных новостных сообщений, взятых из новостных сайтов сети Интернет. 18 текстов относятся к теме *Terrorisme en France* (сюжеты *Explosions*, *Mesures contre terrorisme*, *Suspects*, *Victimes*) и 15 текстов – к теме *Robots* (сюжеты *Robot cuisinier*, *Robot nettoyeur*, *Robot vendeur*, *Robot Nao*).

С целью автоматического определения темы каждого текста была создана первая часть лингвистической базы данных, содержащая словарь дескрипторов, относящих текст к конкретной тематической рубрике. Все дескрипторы были взяты из материала исследования. В табл. 1 приводится фрагмент этой части базы данных.

Таблица 1

Фрагмент первой части лингвистической базы данных

Тип дескриптора	Слово и словосочетание тематической рубрики	
	Terrorisme en France	Robots
Безусловный	*attaque terroriste/attaques terroristes* *attentat terroriste/attentats terroristes* *acte terroriste/actes terroristes* *ceinture explosive/ceintures explosives* *menace terroriste/menaces terroristes* …	*gadget/gadgets* *humanoïde/humanoïdes* *pepper* *robotics* *robotisation* …
Условный квазиоднозначный	*anti-criminalité* *assaillant/assaillants* *assassinat/assassinats* *assaut/assauts* *attaque/attaques* …	*équipement/équipements* *high-tech* *invention/inventions* *innovation/innovations* *machine/machines* …

Для автоматической сегментации ключевых предложений франкоязычных новостных сообщений с целью выделения из них формальных групп и формирования на их основе информационного портрета текста была создана вторая часть лингвистической базы данных. Она включает словарь, единицам

которого приписаны определенные лексико-грамматические теги, указывающие на такие грамматические категории, как часть речи, число, род, время, степень сравнения и т.д. В табл. 2 приводится фрагмент этой части базы данных.

Таблица 2

Фрагмент второй части лингвистической базы данных

№ п/п	Словоформа	Комментарий
1	ACTIONS__N22	N – имя существительное, 2 – женский род, 2 – множественное число.
2	AINSI__D	D – наречие.
3	CETTE__A21	A – имя прилагательное, 2 – женский род, 1 – единственное число.
4	DANS__G	G – предлог.
5	ENCORE__D	D – наречие.
6	ENVOYEZ__V322	V – глагол, 3 – не вспомогательный глагол, 2 – 2-е лицо, 2 – множественное число.
7 ...	ESPRIT__N11 ...	N – имя существительное, 1 – мужской род, 1 – единственное число. ...

Рассмотрим основные особенности процесса решения задачи, связанной с определением тематики текстов новостных сообщений. В память компьютера вводится очередной текст, по которому компьютер составляет список всех слов. Далее компьютер сравнивает каждое слово текста с безусловными дескрипторами лингвистической базы данных по тематике *Terrorisme en France*. Если они найдены, то компьютер ищет далее условные квазиоднозначные дескрипторы тематики *Robots*. В случае, когда условные квазиоднозначные дескрипторы найдены, компьютер делает вывод, что новостной текст политематичен. Если квазиоднозначные дескрипторы тематики *Robots* отсутствуют, то компьютер делает вывод, что новостное сообщение относится к тематике *Terrorisme en France*. Аналогичные действия осуществляются компьютером относительно тематики *Robots*.

Вторая часть формальной модели связана с построением информационных портретов отобранных для исследования текстов франкоязычных новостных сообщений и объединением текстов на основе совпадения их основного содержания в единые сюжеты. Рассмотрим основные особенности данного алгоритма. В памяти компьютера находится текст франкоязычного новостного сообщения, тематическая рубрика которого определена в результате работы описанной выше первой части формальной модели. Как говорилось ранее, для адекватного формализованного представления основного содержания текста необходимо оперировать определенным набором ключевых словосочетаний, которые формируют его информационный

портрет. Процедура выделения ключевых единиц текста базируется на следующих формальных критериях: наибольшей частоте употребления слова в тексте (включая все контекстуальные синонимы и местоименные замены) и максимальном числе абзацев, в которых оно встретилось. Такие слова являются своеобразным центром, вокруг которого формируются другие менее значимые лексические единицы, описывающие определенные микроситуации.

Главные опорные слова каждого текста являются основой для создания его информационного портрета. Построение содержательного портрета текста связано с автоматической сегментацией его предложений и выделением в них формальных групп подлежащего (субъекта), сказуемого (предиката), дополнения (объекта), обстоятельства места (места действия) и обстоятельства времени (времени действия). Эта процедура состоит из двух этапов. На первом этапе проводится тегирование слов текста, то есть приписывание им определенных лексико-грамматических тегов (признаков) с опорой на словарь, содержащийся во второй части лингвистической базы данных. Затем выполняется разрешение анафористичности слов текста. При этом анафористические местоимения (местоимения, которые ссылаются на эксплицитно выраженные именные группы) на основе словарей (абстрактных понятий, географических названий, фамилий, мужских и женских имен и т.д.) заменяются их антецедентами (или референтами). Далее с учетом прямого порядка слов во французском языке и лексико-грамматических тегов осуществляется предварительная сегментация каждого предложения текста. На следующем этапе компьютер проводит анализ полученных сегментов с целью выделения среди них формальных групп подлежащего, сказуемого, дополнения, обстоятельства места и обстоятельства времени. Основная трудность на этом этапе заключается в выделении формальных групп дополнения, обстоятельства места и обстоятельства времени. В результате лингвистического анализа левой и правой контактной дистрибуции имен существительных всех отобранных для исследования текстов франкоязычных новостей были получены списки маркеров левой и правой границ именной группы. К ним относятся единичные словоформы, общие для левой и правой границы (смешанные границы); бинарные сочетания, общие для левой и правой границы (смешанные границы); единичные словоформы, характерные только для левой (или правой) границы; бинарные сочетания, характерные только для левой (или правой) границы.

На заключительном этапе создания информационного портрета компьютер проверяет выделенные ранее формальные группы подлежащего, сказуемого и дополнения на наличие в них главных опорных слов текста. В случае присутствия хотя бы одного главного опорного слова в каждом из трех вышеуказанных сегментов группа подлежащего заносится в поле «субъект», группа сказуемого – в поле «предикат», группа дополнения – в поле «объект», группа обстоятельства места – в поле «место действия», группа

обстоятельства времени – в поле «время действия» информационного портрета текста. После построения информационного портрета текста франкоязычного новостного сообщения последний сравнивается с информационными портретами проанализированных ранее текстов новостей. Сравнение осуществляется по лексическим единицам, входящим в поля «субъект» (подлежащее), «предикат» (сказуемое) и «объект» (дополнение), являющимися ключевыми формальными группами предложения. Если в указанных полях информационных портретов двух текстов есть хотя бы одно совпадение, тексты будут относиться к одному сюжету. Таким образом, по результатам сравнения компьютер определяет, к какому сюжету относится каждое новостное сообщение.

На основе полной формальной модели системы автоматического объединения франкоязычных новостных текстов в единый сюжет на языке программирования *C#* было создано веб-приложение. Компьютерное приложение работает следующим образом. Пользователь указывает путь к папке с файлами в формате *xml*, в которых содержатся разделенные по темам тексты для дальнейшего анализа. Содержание текстового документа выводится на экран. В результате сложной обработки каждого текста компьютер относит его к определенному сюжету. По окончании работы компьютерного приложения на экране демонстрируется итоговое сообщение о результатах обработки всего массива франкоязычных новостных сообщений.

Компьютерный эксперимент по автоматическому объединению 33 текстов франкоязычных новостных сообщений в единые сюжеты показал, что во всех случаях компьютер правильно определил тематику текста и в подавляющем большинстве случаев (93 %) верно объединил тексты в разные сюжеты. Анализ результатов работы компьютерного приложения позволил выявить следующие проблемы:

1. Компьютер проводит безошибочную обработку простых предложений текстов новостей. Однако при синтаксической сегментации ряда предложений возможны ошибки в таких случаях:

а) когда подлежащее слишком отдалено от сказуемого, например, подлежащее уточняется с помощью приложения или в случаях типа *la mise en oeuvre de ce dispositif coutera*;

б) когда сказуемое состоит из нескольких слов, например, *ont été pris en charge*;

в) при обратном порядке слов в вопросительном предложении, например, *Y aura-t-il bientôt un robot dans toutes les gares?*;

г) при обратном порядке слов в косвенной речи, например, *"Les robots Pepper pourront conseiller les clients sur les différents articles proposés en fonction de leurs besoins et goûts", prédit Kohzoh Takaoka, PDG de Nestlé Japon.*

2. Для более совершенной работы программы необходимо составить список устойчивых словосочетаний, которые бы автоматически заносились в определенное поле информационного портрета текста.

3. Добавление к процедуре синтаксического анализа предложений текста новостного сообщения процедуры анализа их семантической структуры повысит точность распределения лексических единиц в полях информационного портрета текста.

Возможность применения полученных результатов видится в том, что представленная в работе формальная модель может стать основой для создания промышленной системы автоматического объединения текстов новостных сообщений в единый сюжет, которую можно использовать в информационных агентствах, а также на радио и телевидении для формирования выпуска новостей.

The article deals with the automatic procedure of French news messages unified in a common plot. This procedure is based on common features of information portraits of such texts: main key words and formal groups of subject, predicate and object extracted from key sentences.

А. Л. Холод
Минск, МГЛУ

ФЛЕЙМ КАК ДЕСТРУКТИВНАЯ СТРАТЕГИЯ ВИРТУАЛЬНОГО ОБЩЕНИЯ

Рассматриваются общие характеристики флейма как деструктивной стратегии интернет-коммуникации. Отмечается, что виртуальная среда обладает целым рядом специфических черт, которые способствуют возникновению речевой агрессии. Для ее выражения коммуниканты часто прибегают к деструктивным стратегиям и тактикам. Одной из таких стратегий является флейм – коммуникативное явление, возникающее в ходе обмена сообщениями в местах многопользовательского сетевого общения, например, на форумах, в чатах, социальных сетях и т.д. Чаще всго флейм реализуется посредством таких коммуникативных тактик, как прямое оскорбление собеседника; оценка его коммуникативных способностей; ироничная оценка сообщений.

Многие исследователи виртуальной среды отмечают, что она является более агрессивной, чем реальный мир [1; 2; 3; 4; 5]. Это обусловлено целым рядом факторов: анонимностью электронного общения; позиционированием унифицированного пользователя Интернета как равноправного члена интернет-сообщества (без учета возрастных, гендерных, расовых, социальных и других особенностей); повышенной эмоциональностью общения; наличием ярко выраженной игровой составляющей интернет-коммуникации; отсутствием формального этикета; более легким, по сравнению с реальностью, установлением контакта; доступностью практически любых каналов общения.

Для выражения речевой агрессии по отношению к членам сетевого сообщества коммуниканты нередко прибегают к деструктивным стратегиям и тактикам, направленным на сознательное и преднамеренное нанесение своему собеседнику морального вреда, испытывая при этом чувство