Синтез распознавателей семантических объектов

Ю.М. Вишняков, Р.Ю. Вишняков

***Аннотация*— Расширение сферы обработки естественно-языковой информации (NLP) и появление в этой связи новых задач вызвали возрастающий интерес как к формализации непосредственно естественно-языковых процессов, так и созданию формализованных инструментов разработки. Одним из таких направлений можно считать поиск и выявление лингвистически присутствующих в текстовых потоках определенных объектов, реализующих свои цели и намерения. Предлагаемая работа выполнена в рамках данного направления и развивает формальный метод проектирования распознавателя для идентификации семантических объектов в естественно-языковых текстовых потоках по оставляемым ими лингвистическим следам. В рамках исследования разработана формальная модель семантического объекта, включающая такие понятия как функция поведения, сценарий, лингвистический след. Предложена формальная модель распознавателя семантических объектов и функция распознавания. разработана формальная методика синтеза распознавателя семантических объектов, основанная на алгебре регулярных выражений и автоматной модели в виде системы переходов. Для сравнения текстовых фрагментов на семантическую близость использовано разработанное ранее вычислительное представление смысла. Предложенные решения в первую очередь рекомендуются для обнаружения и предотвращения преступлений в социальных сетях.**

***Ключевые слова* — семантическая близость, семантический объект, семантическое сравнение, семантический след, семантическое распознавание.**

I. Введение

Открытость и массовая доступность интернет-коммуникаций способствовали появлению новых практических задач, одну из которых составляет выявление и идентификация объектов в естественно-языковых текстовых потоках [1]. Под такими потоками понимаются чаты мессенджеров, социальных сетей и другие документные потоки, а под объектами, например, – обсуждаемый художественный фильм, или какое-нибудь хобби, или – преступление и пр.

Статья получена 27 октября 2024.

Исследование выполнено при финансовой поддержке Кубанского научного фонда в рамках научно-инновационного проекта «НИП-20.1.4.

Юрий Муссович Вишняков, профессор факультета компьютерных технологий и прикладной математики, Кубанский государственный университет (e-mail: jury.vishnyakov@gmail.com)

Ренат Юрьевич Вишняков, доцент факультета математики и компьютерных наук, Кубанский государственный университет (e-mail: jury.vishnyakov@gmail.com)

Укажем, наверное, на самую главный область, которая актуализируют подобного рода исследования – это участившиеся случаи совершаемых в киберпространстве преступлений, и, в особенности, в социальных сетях. В обществе уже сформировался запрос на создание адекватных и эффективных мер противодействия подобного рода преступлениям. На сегодня рост числа киберпреступлений настолько велик, что они уже могут нанести невосполнимый урон государству и обществу.

Например, стоит обратить внимание на публикацию обозревателя «Новой газеты» Г. Мурсалиевой (НГ № 51 от 16.05.2016г.), которая рассматривалась в нашей работе [3]. В публикации шла речь об обосновавшейся в соцсети ВКонтакте преступной организации «Синий кит», склоняющей подростков к суициду. Владение приемами психологического воздействия и возможностями социальных сетей для сокрытия следов преступлений позволяло преступникам долгое время действовать безнаказанно, создавая угрозы обществу, о чем свидетельствует большое число жертв преступлений. Очевидно, что подобные преступные деяния не могут оставаться безнаказанными.

Однако выявление подобного рода преступлений и преступных деяний, наталкивается на серьезные трудности, так как преступники присутствуют в социальных сетях виртуально и лингвистически, используя всячески их возможности для сокрытия следов своих преступлений.

И, тем не менее, такими инструментами противодействия могли бы быть различного рода распознаватели и идентификаторы, способные автоматически обрабатывать естественно-языковую информацию, выделять в ней специфические смысловые черты преступных деяний, распознавать и идентифицировать их по характерным чертам.

Учитывая важность данной проблемы, исследованиями социальных данных занимаются такие университеты как Карнеги-Меллона, Стэнфордский, Оксфордский, INRIA и др., а также крупнейшие мировые корпорации Google, Yhoo!, LinkedIn и пр. Компании-владельцы сервисов социальных сетей активно инвестируют разработку проектов Cassandra, Presto, FlockDB, Thrift и др., которые предназначены для обработки больших массивов пользовательских данных. Успешно развиваются проекты по доступу к хранилищам социальных данных (GNIP), их сбору по заданным сценариям (80legs), социальной аналитике (DataSift), а также расширению существующих платформ с помощью социальных данных (FlipTop). В научные исследования социальных технологий

вливаются огромные суммы денег не только корпорациями, но и государственными ведомствами, в том числе и военного назначения. Идентификация угроз и преступлений в киберпространстве относится к важнейшим проблемам и требует наискорейшего эффективного решения.

Возвращаясь к понятию объекта, отметим, что он обладает индивидуальностью и отображается лингвистически в некоторый образ. Этот образ будем называть семантическим объектом. Семантический объект представлен формализованным описанием, включающим цель, поведение и лингвистические характеристики.

Находясь в текстовом потоке, семантический объект оставляет в нем последовательности характерных лингвистических выражений (семантических следов), по которым можно судить о его присутствии в данном текстовом потоке.

Основная цель предлагаемого исследования состоит в том, чтобы по заранее известному описанию семантического объекта и оставляемым лингвистическим следам его распознать и идентифицировать в текстовом потоке.

Подобная задача обсуждалась нами в работах [1, 2], а предлагаемая работа уточняет и развивает данные исследования, а также иллюстрирует прикладное использование вычислительной теории семантической интерпретации [3,4].

II. Проблема и метод решения

A. *Формальные определения.*

Текстовый поток T это множество упорядоченных во времени и различимых по смыслу естественно-языковых цепочек – токенов (слов, предложений и/или их фрагментов):

$$T = \{t_1, t_2, \ldots, t_m\}, \quad (1)$$

где t_i отдельный токен, являющийся единицей текстового потока. В дальнейшем для упрощения рассуждений будем исходить из того, что токен это предложение.

Семантический объект *SemObj* (Semantic objekct) представляется тройка вида:

$$SemObj = (Q, P, Z), \quad (2)$$

где *Q* –множество лингвистических характеристик, *P* – функция поведения, *Z* –цель.

Лингвистическая характеристика $q_i \in Q$ представляет собой слова, целостные по смыслу текстовые фрагменты или предложения, которые используются для достижения цели *Z*. Множество *Q*, по сути, есть ничто иное как словарь лингвистических заготовок, из которых собирается направленный на достижение цели *Z* поведенческий сценарий. В общем случае существует множество *Sc* (Scenario) таких сценариев, а сам сценарий представляет собой последовательность $\alpha \in Sc$ лингвистических характеристик, причем $Sc \subset Q^*$, где Q^* есть итерация множества Q.

Функцию поведения *P* представим функциональным соответствием (функцией) вида $P: Q^* \rightarrow \{0,1\}$ на множестве цепочек Q^* и множестве {0,1}. Если цепочка $\beta \in Sc$, то $P(\beta)=1$, в противном случае $P(\beta)=0$.

В основу формальной поведенческой модели SemObj можно положить различные подходы.

В нейросетевом подходе требуется формирования для множества сценариев *Sc* обучающей выборки и функции ошибки, а последующего проведения обучения. В случае бесконечного множества *Sc* возможен вопрос о доверии к результатам, что связано с репрезентативностью обучающей выборки. Ответы на все это предполагают далеко не тривиальные процедуры и действия.

В формально грамматическом подходе требуется построить некоторую грамматику G[Z], порождающую множество сценариев Sc. В грамматике цель Z является начальным символом, множество лингвистических характеристик Q образуют словарь терминальных символов, отдельный сценарий $\propto$ является предложением, а их множество образует язык L(G[Z]) грамматики. Достижения цели моделируется поиском вывода $Z \Rightarrow + \alpha$.

В подходе на основе конечных автоматов в нотации алгебры регулярных выражений [5] множество сценариев Sc представляется множеством значений некоторого регулярного выражения E, которое изначально должно быть известно или построено по формализованному описанию SemObj. Далее по E синтезируется *SemObj* в виде модели конечного автомата [5], который, получая на вход любой из сценариев α, последовательно его отрабатывает (разбирает, распознает). Обсудим данный подход, поскольку он наиболее прост в программной реализации и, как нам представляется, покрывает большинство практических задач. Кроме того, в языке программирования Pyhton имеется средства описания и обработки подобного типа данных.

B. *Конструирование (синтез) семантического объекта в виде приведенной системы переходов*

Обобщенное поведение семантического объекта в модели системы переходов представляется диаграммой состояний (Рис. 1). Она имеет одно начальное состояние *S*, одно заключительное – *Z* и дугу *E*, ведущую из состояния *S* в состояние *Z*. Такая диаграмма называется системой переходов.

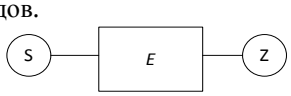

Рис. 1: Система переходов семантического объекта

Дуга *E* имеет следующий смысл - любая цепочка $\alpha_i \in Sc$ переводит SemObj из состояния S в состояние Z, тем самым реализуя разбор сценария.

Для нашего случая алгебра регулярных выражений выглядит следующим образом.

a) $\emptyset$ (пустое множество), λ (пустая цепочка) и лингвистические характеристики $q_1, q_2, \ldots, q_n$ есть регулярные выражения (аксиома);
b) операции над регулярными выражениями (символы операций закавычены):
   - операция ИЛИ – "|";

- операция И – “*” (обычно в символ операции в выражениях опускают;
- операция итерации “{}”
- операция ().

c) если e, e_1 и e_2 – регулярные выражения, то (e), {e}, $e_1|e_2$ и e_1e_2 (символ операции * опущен) также являются регулярными выражениями и других регулярных выражений нет.

Определению алгебры соответствуют представленные на Рис. 2 элементарные конструкции систем переходов:

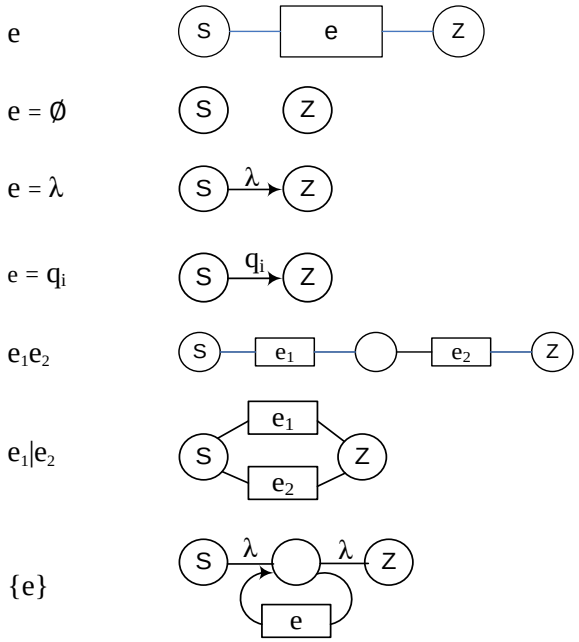

Рис. 2: Конструкции системы переходов для элементарных регулярных выражений

Процесс синтеза *SemObj* сводится к пошаговой декомпозиции системы переходов (Рис. 1) по правилам (Рис.2) до тех пор, пока все дуги не будут взвешены лингвистическими характеристиками и символом λ. На этом процесс декомпозиции завершается, а полученная система переходов называется приведенной.

Например, для регулярного выражения $\{(q_1|q_2)\}q_3$ пошаговый процесс декомпозиции *SemObj* показан на Рис. 3.

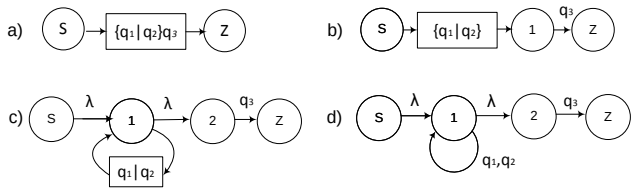

Рис. 3: Процедура декомпозиции системы переходов

SemObj выполняет разбор сценария последовательно слева направо по лингвистическим характеристикам.

Например, если для *SemObj* (Рис. 3d) задать следующие сценарии:

a) $\alpha_1 = q_3$;
b) $\alpha_2 = q_1q_2q_3$;
c) $\alpha_2 = q_1q_1q_2q_2q_3$,

то последовательности смены состояний для каждого из сценариев приведены на Рис. 4 и представляют они собой путь из состояния S в состояние Z в диаграмме состояний приведенной системе переходов и называются линейными развертками конечного автомата.

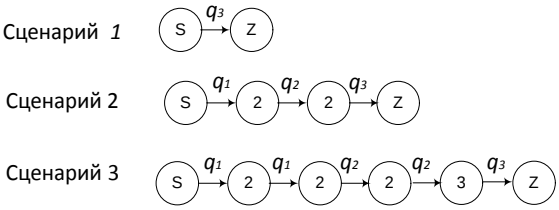

Рис. 4: Линейные развертки *SemObj*

По сути, линейная развертка представляет собой отображение приведенной системы переходов на сценарий.

C. Распознавание семантических объектов

Пусть семантический объект SemObj задан множеством поведенческих сценариев Sc и приведенной системой переходов, пусть также задан текстовый поток T. Требуется построить процедуру поиска и распознавания в текстовом потоке T семантического объекта SemObj и оценить степень доверия к результатам данной процедуры.

Предположим, что SemObj реализует сценарий $\alpha=q_{r1}q_{r2}\ldots q_{rn}$, где $\alpha\in Sc$, $q_{ri}\in Q$. Данный сценарий в текстовом потоке отображается в текстовый поток T последовательностью $\ldots t_{r1}\ldots t_{r2}\ldots t_{rn}\ldots$ токенов. Удалив из нее незначащие токены, получим нормализованную последовательность $\alpha'=t_{r1}t_{r2}\ldots t_{rn}$ токенов.

Последовательность α' представляет результат отображения сценария α на текстовый поток, или, говоря иначе, ее можно рассматривать как след α' сценария α в текстовом потоке.

Введем функцию распознавания вида $F(\alpha,\alpha')$, принимающую значения на интервале [0..1].

Если F=0, то семантический объект в текстовом потоке не обнаружен, если F=1 – полностью распознан и идентифицирован.

Сконструируем функцию распознавания, для чего нам понадобится функция сравнения двух текстовых отрезков на семантическое подобие (близость) [3,4]. Если имеются два текстовых фрагментов a и b со смысловыми значениями $S(a)$ и $S(b)$ (пока для нас важен пока только факт существования этих значений) соответственно, то функция вида $Cprox(S(a), S(b))\in[0..1]$ определяет их смысловую близость. Если $Cprox = 0$, то семантическая близость между a и b отсутствует, при $Cprox = 1$ имеет место полная семантическое совпадение.

Пусть длина последовательности $|\alpha|$ равна длине последовательности $|\alpha'|$ и между ними имеется полное поэлементное взаимно-однозначное соответствие лингвистических характеристик сценария α и токенов последовательности α'.

Рассмотрим тройку вида (q_i, t_{rj}, w_i), где q_i и t_{rj} это пара из соответствия, а $w_I = Cprox(q_i, t_{ri})$ есть мера

семантической близости q_i и t_{rj}. Данную меру можно рассматривать как меру присутствия характеристики q_i в текстовом потоке в виде токена-следа t_{rj}. Также данное высказывание можно определить как $t_{ri} = (q_i, w_i)$, тогда с учетом данного факта последовательность α' (семантический след сценария α) можно представить в виде:

$$\alpha' = (q_1, w_1)\ (q_2, w_2)\ \dots\ (q_т, w_n) \quad (3)$$

Очевидно, что функцию распознавания F можно представить средневзвешенной величиной вида:

$$F(\alpha,\alpha') = (w_1+w_2+\dots+w_n)/n\ , \quad (4)$$

и она удовлетворяет всем сформированным требованиям к функции распознавания. Следует отметить, что для функции распознавания можно подбирать и другие представления, исходя из решаемых практических задач.

Случаи по формированию функции распознавания, когда в текстовом потоке обнаруживаются семантические следы фрагментов сценариев, требуют несколько отдельного обсуждения. По этой причине в настоящую работу данное обсуждение не включено.

И так суть идентификации семантического объекта по его лингвистическому описанию сводится к последовательному сканированию токенов естественно-языкового текстового потока автоматной моделью SemObj и их сравнению на семантическую близость с лингвистическими характеристиками текстового потока.

D. Практические результаты

Предложенные теоретические наработки представляют основу учебного комплексного проекта, предлагаемого команде студентов разных курсов в виде заданий по научно-исследовательской работе, курсовому и дипломному проектированию, он постоянно совершенствуется и дополняется.

Проект непосредственно включает следующие функциональные подсистемы:

- подсистему подготовки и нормализации текстового потока;
- подсистему распознавания семантического объекта;
- подсистему семантического сравнения токенов лингвистических характеристик;
- подсистему формального конструирования и редактирования семантического объекта;
- подсистему отладки и журналирования.

Работа комплекса сводится к следующим этапам. На этапе подготовки комплекса существует некоторая исходная описательная информации об объекте, присутствие которого нужно выявить в текстовом потоке. Предполагается, что данная информация получена от внешнего пользователя. На основе этой описательной информации в подсистеме конструирования и редактирования семантического объекта готовится его формальное описание и модель. Создается множество Q его лингвистических характеристик и формируется поведенческая модель E. Формальное описание семантического объекта сохраняется, чтобы его в последующем можно было при необходимости редактировать.

Внешний пользователь также предоставляет текстовый поток, по которому просит обеспечить поиск семантического объекта. Данный текстовый поток подлежит обработке подсистемой подготовки и нормализации. Текстовый поток может очищаться путем удаления стоп-слов, токенизироваться для разбиения текста на предложения и слова и пр. Далее путем удаления информационно незначащих токенов выполняется нормализация текстового потока.

Следующий этап представляет непосредственно работа комплекса. Нормализованный текстовый поток подается на вход подсистемы распознавания семантического объекта. Данная подсистема, пошагово сканируя текстовый поток, сравнивает каждый из токенов на семантическую близость со списком лингвистическими характеристиками, выбирая из списка наиболее подходящую из характеристик. После этого распознаватель переводится в новое состояние в соответствие с диаграммой состояний. Для семантического сравнения на семантическую близость токен и лингвистическая характеристика передаются подсистеме семантического сравнения, которая, семантически сравнивая их, выдает распознавателю меру семантического сходства.

Продвигаясь подобным образом по текстовому потоку, распознаватель восстанавливает сценарий работы семантического объекта и формирует функцию распознавания. По завершении обработки по значению функции распознавания можно уже судить о присутствии семантического объекта в текстовом потоке с оценкой степени доверия.

Подсистема отладки и журналирования предназначена, как это следует из названия, для отладки работы комплекса и хранения результатов обработки.

В качестве примера ниже приведены фрагменты правдоподобного текстового потока и результаты семантического сравнения токенов с лингвистическими характеристиками семантического объекта. Этот фрагмент текстового потока отражает диалог, который возможно мог бы состояться реально у семантического объекта «Синий кит» с его жертвой.

21.	(q_{15},1,0)	К. Мы закалим твой характер.
22.	(q_{16},1,0)	К. Проблемы сами уйдут.
23.		К. Нашем?
24.		Ж. Да, хорошо, попробуем.
25.	(q_{17},1,0)	К. Это твое первое задание.
26.	(q_{18},1,0)	К. Сядь на стул и замри на 15 минут.
27.		Ж. У меня получилось! Ура!
28.	(q_{19},1,0)	К. Молодец, ты справился!
29.	(q_{15},0,5)	К. Завтра новое задание

53.		Ж. Родители любят меня!
54.	(q_{28},1,0)	К. Ты никому не нужен!
55.	(q_{24},1,0)	К. Сделай ЭТО!
56.	(q_{26},0,7)	К. Избавься от страданий!
57.	(q_{24},1,0)	К. Сделай ЭТО
58.		Ж. Я боюсь, пожалей меня!
59.	(q_{29},1,0)	К. Не ты меня, наказание будет страшным
60.	(q_{24},1,0)	К. Сделай ЭТО!

Рис. 5: Фрагменты примера текстового потока

Ряд теоретических наработок, таких как семантическое сравнение, модель семантического объекта реализовывались на языке Pyhton и показали свою работоспособность. Для построения синтаксических деревьев в подсистеме семантического сравнения использовались библиотеки NLTK, Natasha (набор Python-библиотек для обработки текстов на естественном русском языке), Deep Pavlov в доступных

версиях. В подсистеме нормализации текстов использованы библиотека NLTK.

III. Заключение

В работе предложен формальный метод синтеза распознавателя для идентификации семантических объектов в естественно-языковых текстовых потоках по оставляемым ими лингвистическим следам. Разработана формальная модель семантического объекта, функция поведения, сценарий, лингвистический след. Предложена формальная модель распознавателя семантических объектов, функция распознавания и методика синтеза распознавателя, основанная на алгебре регулярных выражений и автоматной модели в виде системы переходов. Для сравнения текстовых фрагментов на семантическую близость использовано ранее разработанное вычислительное представление смысла. По мнению авторов предложенные решения в первую очередь можно рекомендовать для обнаружения и предотвращения преступлений в социальных сетях.

Авторы посчитали также возможным дать ссылки на фундаментальные работы и работы, которые в той или иной мере касаются проблем обработки естественно-языковой информации [6-25], хотя перечень подобного рода работ приведенным списком не исчерпывается.

Кроме того, в работе приведены некоторые сведения о программных решениях, разработанных по данным исследованиям

Благодарности

Исследование выполнено при финансовой поддержке Кубанского научного фонда в рамках научно-инновационного проекта «НИП-20.1.4»

Библиография

[1] Y.M. Vishnyakov und R.Y. Vishnyakov Identification of semantic objects in information stream Journal of Physics: Conference Series; Bristol Том 1902, Изд. 1, (May 2021). DOI:10.1088/1742-6596/1902/1/012104.

[2] Вишняков Ю.М., Вишняков Р.Ю. Формализация распознавания и идентификации семантических объектов в естественно-языковых текстовых потоках// Известия ЮФУ. Технические науки, 2024, №4 – С.110-128

[3] Yury M. Vishnyakov, Renat Yu. Vishnyakov The Linguistic Proximity in Information Retrieval and Document Classification. // 14th IEEE International Symposium on Computational Intelligence and Informatics to be held on November 19-21, 2013 in Budapest, Hungary. p. 131-134.

[4] Yuri M. Vishnyakov, Renat Y. Vishnyakov Computational theory of semantics representation in scientific and technical texts // AMCSM_2018 IOP Publishing IOP, Conf. Series: Journal of Physics: Conf. Series 1202 (2019) 012008 doi:10.1088/1742-6596/1202/1/012008.

[5] Philip M. Lewis, Daniel J. Rosenkrantz, Richard E. Stearns, R. E. Stearns Compiler Design Theory/ Addison-Wesley Publishing Company, 1976, 647 s.

[6] Koncel-Kedziorski R, Hajishirzi H and Sabharwal A et Al. 2015 Parsing algebraic word problemsinto equations Transactions of the Association for Computational Linguistics, 3:585–597.

[7] Devlin J, Chang MW, Lee K and Toutanova K 2018 Bert: Pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv:1810.04805

[8] Cruse A 2011 Meaning in language: An introduction to semantics and pragmatics Oxford University Press UK.

[9] Налимов В.В. Вероятностная модель языка. О соотношении естественных и искусственных языков. М.: Наука, 1979, 303 с.

[10] Николаев И.С. Компьютерная и прикладная лингвистика / Николаев И.С., Митренина О.В., Ландо Т.М. (ред.) – М.: Ленанд, 2016. – 316 с.

[11] Тестелец Я.Г. Введение в общий синтаксис. Учебное пособие. М.: Изд-во Российского гуманитарного университета, 2001, - 830 с.

[12] Прохоренок Н.А., Дронов В.А. Python 3. Самое необходимое. СПб.: БХВ-Петербург, 2019. – 608 с.

[13] Бенгфорт Бенджамин Прикладной анализ текстовых данных на Python / Машинное обучение и создание приложений обработки естественного языка / Бенгфорт Бенджамин, Билбро Ребекка, Охеда Тони — СПб.: Питер, 2019. — 368 с.

[14] Devlin J, Chang MW, Lee K and Toutanova K Bert: Pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv:1810.04805, 2018

[15] Hu K, Wu H and Qi K et Al. A domain keyword analysis approach extending Term Frequency- Keyword Active Index with Google Word2Vec model Scientometrics, Springer, 2017, 1 –38.

[16] Tianshuo Peng, Zuchao Li, Lefei Zhang, Hai Zhao, Ping Wang, Bo Du; Multi-modal Auto-regressive Modeling via Visual Tokens, MM '24: Proceedings of the 32nd ACM International Conference on Multimedia; https://doi.org/10.1145/3664647.3681685

[17] Aman Bhadouria, Pranav Gupta, Parish Bindal, Kapil Madan, Sonal Sonal; Automated Examination System using Machine Learning and Natural Language Processing, IC3-2024: Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing; https://doi.org/10.1145/3675888.3676144

[18] Azhar Kassem Flayeh, Yaser Issam Hamodi, Nashwan Dheyaa ZakiText Analysis Based on Natural Language Processing (NLP), 2022 2nd International Conference on Advances in Engineering Science and Technology (AEST); DOI: 10.1109/AEST55805.2022.10413039

[19] Xin Wu, Yi Cai, Zetao Lian, Ho-fung Leung, Tao Wang; Generating Natural Language From Logic Expressions With Structural Representation, IEEE/ACM Transactions on Audio, Speech, and Language Processing (Volume: 31); DOI: 10.1109/TASLP.2023.3263784

[20] Lalitha Manasa Chandrapati, Ch. Koteswara Rao; Descriptive Answers Evaluation Using Natural Language Processing Approaches, IEEE Access (Volume: 12) 21 June 2024; DOI:10.1109/ACCESS.2024.3417706

[21] Komal Kalra, Raj Gaurang Tiwari; Exploring Common Areas and Types of Cybercrime in Today's Digital Landscape, 2023 3rd Asian Conference on Innovation in Technology (ASIANCON); DOI: 10.1109/ASIANCON58793.2023.10270422

[22] Zhenhua Zhao, Chao Wang, Shaopei Ji; Text Similarity Calculation Model Based on Semantic Information and Syntactic Structure Fusion Weighting, 2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE); DOI:10.1109/CISCE62493.2024.10653095

[23] Anshul Modi, Yuvraj Singh Dhanjal, Anamika Larhgotra; Semantic Similarity for Text Comparison between Textual Documents or Sentences, 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES); DOI:10.1109/ICSES60034.2023.10465440

[24] Sonali Mhatre, Shilpa Satre, Mansi Hajare, Aditi Hire, Aniket Itankar, Shruti Patil; Text Comparison Based on Semantic Similarity, 2023 3rd International Conference on Intelligent Technologies (CONIT); DOI: 10.1109/CONIT59222.2023.10205616

[25] Aadeesh Bali, Aniket Bhagwat, Aditya Bhise, Sarang Joshi; Semantic Similarity Detection and Analysis For Text Documents, 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE); https://doi.org/10.1109/ic-ETITE58242.2024.10493834

Synthesis of Semantic Object Recognizers

Yuri Vishnyakov, Renat Vishnyakov

***Abstract* – The expansion of the natural language processing (NLP) domain and the emergence of new tasks in this field have sparked increasing interest in both the formalization of natural language processes and the development of formalized design tools. One such area is the identification and recognition of certain objects that are linguistically present in text streams and that fulfill specific goals and intentions. The present study is conducted within this area and develops a formal method for designing a recognizer to identify semantic objects in natural language text streams based on their linguistic traces. As part of the research, a formal model of a semantic object was developed, which includes concepts such as behavior function, scenario, and linguistic trace. A formal model of a semantic object recognizer and a recognition function are proposed. A formal method for synthesizing the recognizer is developed based on regular expression algebra and an automaton model in the form of a transition system. A previously developed computational representation of meaning was used to compare text fragments for semantic proximity. The proposed solutions are primarily recommended for detecting and preventing crimes in social networks.**

***Keywords* – semantic proximity, semantic object, semantic comparison, linguistic trace, semantic recognizer.**

REFERENCES

[1] Y.M. Vishnyakov, R.Y. Vishnyakov. Identification of semantic objects in information stream. *Journal of Physics: Conference Series; Bristol*. Vol. 1902, 1 (May 2021). doi: 10.1088/1742-6596/1902/1/012104

[2] Yu.M. Vishnyakov, R.Yu. Vishnyakov. Formalization of recognition and identification of semantic objects in natural language text streams. *Izvestiya SFedU. Engineering Sciences*. 2024, No. 4. P. 110-128.

[3] Yu.M. Vishnyakov, R.Yu. Vishnyakov. The Linguistic Proximity in Information Retrieval and Document Classification. In: *14th IEEE International Symposium on Computational Intelligence and Informatics*. Budapest, Hungary; 2013. P. 131-134.

[4] Yuri M. Vishnyakov, Renat Y. Vishnyakov Computational theory of semantics representation in scientific and technical texts. *AMCSM_2018 IOP Publishing IOP, Conf. Series: Journal of Physics: Conf. Series*. 2019. Vol. 1202. 012008. doi: 10.1088/1742-6596/1202/1/012008.

[5] Philip M. Lewis, Daniel J. Rosenkrantz, Richard E. Stearns, R. E. Stearns Compiler Design Theory. Addison-Wesley Publishing Company, 1976, 647 s.

[6] Koncel-Kedziorski R, Hajishirzi H and Sabharwal A et Al. 2015 Parsing algebraic word problemsinto equations. *Transactions of the Association for Computational Linguistic*s, 3:585–597.

[7] Devlin J, Chang MW, Lee K and Toutanova K 2018 Bert: Pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv:1810.04805

[8] Cruse A. Meaning in language: An introduction to semantics and pragmatics Oxford University Press UK, 2011.

[9] Nalimov V.V. Verojatnostnaja model' jazyka. O sootnoshenii estestvennyh i iskusstvennyh jazykov. M.: Nauka, 1979, 303 p.

[10] Nikolaev I.S. Komp'juternaja i prikladnaja lingvistika / Nikolaev I.S., Mitrenina O.V., Lando T.M. (eds.) M.: Lenand, 2016. 316 p.

[11] Testelec Ja.G. Vvedenie v obshhij sintaksis. Uchebnoe posobie. M.: Izd-vo Rossijskogo gumanitarnogo universiteta, 2001. 830 p.

[12] Prohorenok N.A., Dronov V.A. Python 3. Samoe neobhodimoe. SPb.: BHV-Peterburg, 2019. 608 p.

[13] Bengfort Bendzhamin Prikladnoj analiz tekstovyh dannyh na Python / Mashinnoe obuchenie i sozdanie prilozhenij obrabotki estestvennogo jazyka / Bengfort Bendzhamin, Bilbro Rebekka, Oheda Toni. SPb.: Piter, 2019. 368 p.

[14] Devlin J, Chang MW, Lee K and Toutanova K Bert: Pre-training of deep bidirectional transformers for language understanding arXiv preprint arXiv:1810.04805, 2018.

[15] Hu K., Wu H., Qi K. et Al. A domain keyword analysis approach extending Term Frequency- Keyword Active Index with Google Word2Vec model Scientometrics, Springer, 2017, p. 1-38.

[16] Tianshuo Peng, Zuchao Li, Lefei Zhang, Hai Zhao, Ping Wang, Bo Du; Multi-modal Auto-regressive Modeling via Visual Tokens, MM '24. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. 2024. doi: https://doi.org/10.1145/3664647.3681685

[17] Aman Bhadouria, Pranav Gupta, Parish Bindal, Kapil Madan, Sonal Sonal; Automated Examination System using Machine Learning and Natural Language Processing. In: *IC3-2024: Proceedings of the 2024 Sixteenth International Conference on Contemporary Computing*. 2024. doi: https://doi.org/10.1145/3675888.3676144

[18] Azhar Kassem Flayeh, Yaser Issam Hamodi, Nashwan Dheyaa ZakiText Analysis Based on Natural Language Processing (NLP), *2022 2nd International Conference on Advances in Engineering Science and Technology (AEST);* doi: 10.1109/AEST55805.2022.10413039

[19] Xin Wu, Yi Cai, Zetao Lian, Ho-fung Leung, Tao Wang; Generating Natural Language From Logic Expressions With Structural Representation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (Volume: 31); doi: 10.1109/TASLP.2023.3263784

[20] Lalitha Manasa Chandrapati, Ch. Koteswara Rao; Descriptive Answers Evaluation Using Natural Language Processing Approaches, *IEEE Access*. Vol. 12. 2024. doi: 10.1109/ACCESS.2024.3417706

[21] Komal Kalra, Raj Gaurang Tiwari; Exploring Common Areas and Types of Cybercrime in Today's Digital Landscape. In: *2023 3rd Asian Conference on Innovation in Technology (ASIANCON).* doi: 10.1109/ASIANCON58793.2023.10270422

[22] Zhenhua Zhao, Chao Wang, Shaopei Ji; Text Similarity Calculation Model Based on Semantic Information and Syntactic Structure Fusion Weighting. In: *2024 6th International Conference on Communications, Information System and Computer Engineering (CISCE)*. doi: 10.1109/CISCE62493.2024.10653095

[23] Anshul Modi, Yuvraj Singh Dhanjal, Anamika Larhgotra; Semantic Similarity for Text Comparison between Textual Documents or Sentences. In: *2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES).* doi: 10.1109/ICSES60034.2023.10465440

[24] Sonali Mhatre, Shilpa Satre, Mansi Hajare, Aditi Hire, Aniket Itankar, Shruti Patil; Text Comparison Based on Semantic Similarity. In: 2023 3rd International Conference on Intelligent Technologies (CONIT). doi: 10.1109/CONIT59222.2023.10205616

[25] Aadeesh Bali, Aniket Bhagwat, Aditya Bhise, Sarang Joshi; Semantic Similarity Detection and Analysis For Text Documents, 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE); https://doi.org/10.1109/ic-ETITE58242.2024.10493834