УДК 004.738.52

О.С. Волковський, Є.Р. Ковилін

КОМПЬЮТЕРНАЯ СИСТЕМА АВТОМАТИЧЕСКОГО АНАЛИЗА ПРОМЫШЛЕННЫХ ИНСТРУКЦИЙ

Аннотация. На основе анализа существующих подходов к созданию систем автоматической обработки текстов, произведен выбор модели представления текстовой производственной инструкции в формате доступном для программной реализации. Разработана структура прикладной системы выявления участков с низкой смысловой связностью. Проработаны алгоритмические решения для каждого из этапов обработки текстовой инструкции. Осуществлено тестирование системы на основе государственных инструкций по технике безопасности на производстве.

Введение. Процесс составления промышленных инструкций сегодня является довольно актуальной проблемой по нескольким причинам. Во-первых, несмотря на обширное количество ГОСТов и правил, зачастую игнорируется семантическая ценность и четкость составляемых текстов. В случае инструкций по технике безопасности или инструкций касающихся обращения с опасным тяжелым промышленным оборудованием этот фактор может привести к трагичным последствиям. Во-вторых, составление и последующая проверка инструкций производится людьми, из-за чего существует вероятность возникновения ошибок вследствие человеческого фактора. Наконец, текущие стандарты составления инструкций описывают далеко не все аспекты промышленной деятельности и эксплуатации производственного оборудования. Учитывая тенденцию к компьютеризации практически всех направлений человеческой деятельности, автоматизация процесса составления производственных инструкций является наиболее оптимальным решением описанных проблем. С этой целью была разработана программная пользовательская система, снабженная интеллектуальным компонентом автоматической обработки текстов, речь о которой и пойдет в этой статье.

 ISSN 1562-9945

Анализ последних исследований. В научной отрасли современной отечественной АОТ не существует прикладных аналогов разрабатываемой системы. Однако возможно выделить несколько аналогичных классов систем обработки текста, отличающихся между собой и сложностью обработки данных, и сложностью интеллектуального компонента. Условно, данные системы можно разделить по моделям языка, заложенным в их основу, на три типа: генеративные грамматики Хомского, семантическая сеть и инструменты нейронных сетей. Поскольку необходимость в прикладной реализации, в данном случае, стоит выше теоретической составляющей, рассмотрим наиболее современные конкретные программные системы каждого из указанных классов. Для начала, рассмотрим программу генерации тестовых заданий для дистанционного обучения студентов, основывающуюся на парадигмах формальных грамматик Хомского [2]. Генеративная грамматика составляющих отталкивается от аксиомы существования явления языковой компетенции, заключающейся в способности человека к усвоению и пониманию естественной человеческой речи, независимо от языка. Следуя этому, генеративная грамматика ставит перед собой цель смоделировать эту способность в рамках порождения правильных предложений, используя определенный конечный набор правил, алфавит и начальный символ предложения, из которого, с помощью специальных грамматических правил, можно разворачивать схемы построения предложения – непосредственные составляющие. Теоретически, множество непосредственных составляющих ничем не ограничено и является бесконечным, на практике – сам язык, предметная область, рабочий текстовый корпус и возможности ЭВМ существенно уменьшают размер множества непосредственных составляющих.

Широкое распространение в области АОТ получила технология семантических сетей, являющаяся следующим витком развития области обработки текстов. Семантическая сеть представляет собой граф, где в вершинах стоят семантические единицы, а дуги описывают смысловые связи между ними. Под семантическими единицами, зачастую, может пониматься и отдельное слово, и предложение, и даже целый документ. Практическое применение семантической сети к задаче АОТ хорошо проиллюстрировано в работе [5] - системе автоматического консультирования. Разработчики ставят перед собой за-

дачу генерации базы знаний отдельной предметной области для обеспечения диалога с пользователем по соответствующим ей вопросам. Семантическую сеть предлагается использовать для хранения извлекаемых знаний на основе текстового корпуса, представляющего собой наборы заранее заготовленных шаблонных фраз – ответов.

Передовым инструментом, применимым к задаче автоматической генерации текстов, являются прикладные методы искусственного интеллекта – реализация процесса АОТ при помощи нейронных сетей. Из работы [6], где рекуррентная сеть используется для автоматического создания описаний о товарах некоторого интернет магазина, видно, что результаты получаются довольно неоднозначные. Главным плюсом подхода является полная автоматизация процесса генерации текста, высокая степень адаптивности системы и низкие затраты на ее настройку и внедрение. Однако, очевидны некоторые проблемы появления «смыслового мусора». Причина этого заключается в том, что несмотря на кажущееся наличие интеллектуальной обработки, система не хранит знаний о семантике того что описывает и генерирует, отталкиваясь лишь от заранее заданных шаблонов – учителей.

Постановка проблемы. Сама по себе, тема автоматической обработки текста (АОТ), затрагивает большое количество научных вопросов, связанных в первую очередь с проблемами поиска алгоритмических решений для таких продуктов. Разработка прикладных программных систем подразумевает выбор того или иного механизма описания и реализации модели данных, доступной для обработки ЭВМ. Однако естественный язык (ЕЯ) является неформализованной системой, обладающей непостоянностью и неоднородностью собственных правил, из-за чего усложняется математическое и алгоритмическое описания его компонент. Главной же проблемой является описание семантических характеристик текста на уровне алгоритмического представления - поскольку ЕЯ это не просто набор слов, основанный на грамматических составляющих это, в свою очередь, приводит многих разработчиков к необходимости учета семантических связей между отдельными словами, предложениями и даже документами. Поставленная же задача автоматизации проверки производственных инструкций требует не только описания семантических связей в тексте, но и наличия механизма получения и оценки семантического веса,

 ISSN 1562-9945

поэтому важной задачей является обоснованный выбор алгоритмической модели естественного языка. Выбранная модель должна иметь не только теоретическую ценность, но и обладать механизмами создания прикладной программной реализации на базе современных возможностей ЭВМ.

Рассмотрев основные подходы к прикладной реализации АОТ, становится возможным осуществить выбор модели языка, позволяющей описать промышленную инструкцию в доступной для алгоритмического представления форме. Говоря о генеративных грамматиках Хомского можно отметить, что рассмотренная система дистанционного обучения в частности, и модель языка Хомского в целом, не решает проблем программного представления семантики, поскольку главной целью грамматики Хомского является вывод грамматически правильных предложений из некоторого алфавита, используя цепочки глубинного уровня. И если теоретически мы можем выводить бесконечно большое количество таких цепочек, что, собственно и позволяет описать абсолютно любой ЕЯ, то на практике это не представляется возможным – даже если отбросить такое свойство языка как изменчивость, количество цепочек будет хоть и не бесконечно, но, безусловно, огромно. И чем более выражена флексия в языке – тем сложнее будет его описать. Помимо всего вышесказанного, применение генеративных грамматик, изначально рассматривалось для парсинга языка программирования, где глубинные структуры представлены цепочками формального языка, и грамматическая правильность важнее смысловой семантики (контроль за этим полностью лежит на плечах программиста), что противоречит нашей изначальной задаче [7].

Использование нейронных сетей может показаться заманчивой идеей, однако, для любого типа нейронной сети необходим набор нормализованных числовых данных, описывающих как обучающую выборку, так и анализируемый текст инструкции, поэтому такой мощный инструмент ИИ следует рассматривать уже как постобработку математической модели естественного языка.

В нашем случае, наиболее важной характеристикой текста промышленной инструкции является именно смысловые связи между ее элементами, поэтому семантические сети это наиболее подходящий выбор для описания модели ЕЯ, на основе которой и производится

оценка и семантический разбор текста промышленной инструкции. Это в свою очередь поднимает вопрос построения обходных путей необходимости составления эталонной базы знаний, на основе которой будет происходить формирование семантической сети. Подход, использованный в работе [5] формирует замкнутую систему, результаты работы которой не выходят за пределы добавленной в нее базы знаний, тогда как в нашей задаче важнейшими параметрами являются адаптивность и универсальность прикладного применения разработки.

Цель работы. В процессе работы перед нами стоит несколько задач. В первую очередь, необходима разработка адаптивного алгоритма построения семантической сети, требующего минимальный и конечный набор знаний для своей корректной работы. Следующим шагом является разработка методики снятия семантических характеристик сети для последующей классификации и оценки. Последним этапом является реализация подхода к получению семантически слабой единицы исходного текста промышленной инструкции.

Изложение основного материала. Создавая любую систему АОТ, первым этапом становится синтаксический анализ исходного текста. В случае семантической обработки инструкции, важным аспектом является так же и получение частей речи для каждого из слов. Это позволит системе работать только со семантически значимыми элементами текста, исключая служебные части речи из анализа. Для этого, в систему был включен размеченный эталонный текстовый корпус, в котором каждое слово соотносится с определенной частью речи. Для каждой такой пары выделяются маркеры-окончания – две и три последние буквы слова, а также окончание, полученное при помощи алгоритма Портера. Эти данные являются обучающей выборкой для наивного байесовского классификатора, где классами являются части речи. На вход обученной модели передается массив слов из текста исходной инструкции, и если для анализируемого слова часть речи была определена как неинформативная – слово исключается из последующего анализа. Проведенные эксперименты показывают, что точность определения частей речи составляет порядка 89%.

Рабочее множество возможно дополнительно сократить, если провести операцию выделения стемм в тексте. Для этого над каждой

парой значимых слов из текста инструкции проводится отсечение окончаний по алгоритму Портера, после чего рассчитывается величина расстояния Левенштейна. Если длина общей части слов больше, чем расстояние Левенштейна для анализируемой пары слов, то оба слова заменяются на их наибольшую общую часть.

На основе проведенного анализа составляется матрица, столбцы которой соответствуют стеммам, строки – предложениям, а значениями являются количество вхождений отдельной стеммы в каждое предложение. Над составленной таким образом матрицей выполняется операция сингулярного разложения. На основе свойства устойчивости мы можем игнорировать значения левой и правой матриц, соответствующие низким сингулярным величинам. В результате этой операции, для каждого предложения и стеммы остается два нормализированных значения, характеризующих их частотную составляющую. Отображение полученных данных на плоскость изображено на рис.1, где для анализа был выбран текст из главы 6.2. «Внутренне освещение» правил устройства электроустановок Украины от 2014 года (ПУЭУ-2014) [9].

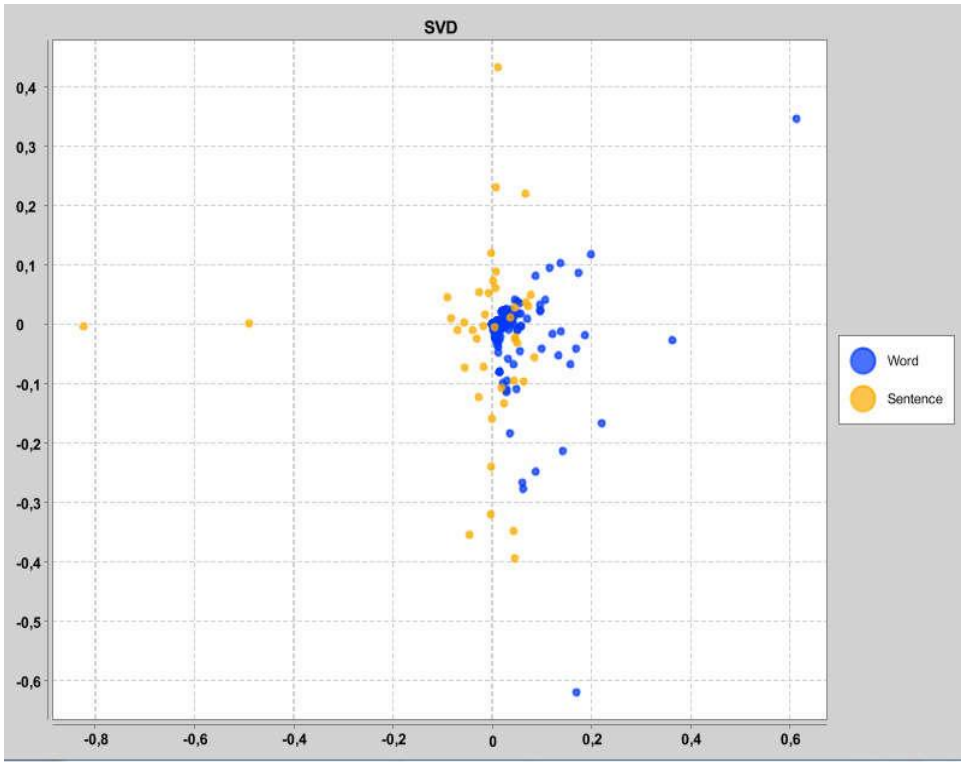

Рисунок 1 – Проекция частотной составляющей инструкции ПУЭУ-2009

После получения частотной числовой картины текста инструкции, необходимо привести полученные данные к некоторой семантической картине, на основе которой возможна генерация семантической сети. Для этого, над полученными данными проводится операция кластеризации по алгоритму k-means, где количество точек-кластеров определяется по формуле (1), в которой count(W) – общее количество слов, (WU)– общее количество уникальных стемм.

$$cl(W, W_U) = \frac{count(W)}{count(W_U)} \quad (1)$$

Значения центроидов кластеров-стемм рассчитываются по формуле (2), где W0...WCL – частотные веса стемм.

$$Cst(W_U) = \max(W_0 \ldots W_{cl}) \quad (2)$$

Значения центроидов предложений рассчитываются по формуле (3), где WS – анализируемое предложение, Wi – вес стеммы в предложении, SN – количество стемм в предложении.

$$Cs(W_S) = \max\left(\sum_{i=0}^{SN} W_i\right) \quad (3)$$

Для каждого кластера – стеммы определяется его вес – количество содержащихся в нем точек, на основе которого формируется каркас будущей семантической сети: кластеры-стеммы связываются между собой в порядке увеличения их веса.

Над значениями каждого кластера-стеммы и кластера-предложения выполняется построение контура выпуклой фигуры по алгоритму Джарвиса. Если контур фигуры-предложения пересекается с контуром фигуры-стеммы, то между ними устанавливается связь, семантический вес которой равен количеству точек, содержащихся в площади пересечения. В результате мы получаем структуру семантической сети, изображенную на рис. 2.

Основываясь на полученных данных, становится возможным получения массива семантически слабых предложений. Для этого достаточно провести анализ кластеров-предложений, не имеющих связи с кластером-стеммой. Если такая семантическая единица была найдена, мы выбираем из нее предложения с наименьшим суммарным весом входящих в них стемм, общее количество которых рассчитывается по формуле 4.

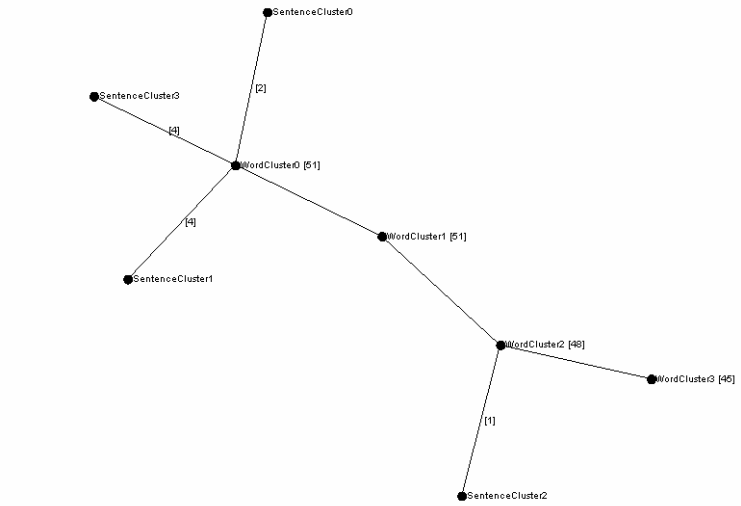

Рисунок 2 – Семантическая сеть инструкции ПУЭУ-2009

$$cl(S_C, S) = round\left(\frac{count(S_C)^2}{count(S)}\right) \quad (4)$$

где count(S) – общее количество предложений, count(SC) – количество предложений в кластере, отношение cl которых описывает такую округленную часть от количества предложений в кластере, которую сам кластер занимает относительно общего количества предложений. Поскольку анализируемая инструкция ПУЭУ-2009 изначально представляет собой семантически связанный текст, для тестирования работы системы добавим в нее предложение из инструкции НПАОП 0.00-1.12-84 Правила взрывобезопасности при использовании мазута и природного газа в котельных установках [10]. Результаты обработки такого текста изображен на рис. 3 и рис. 4.

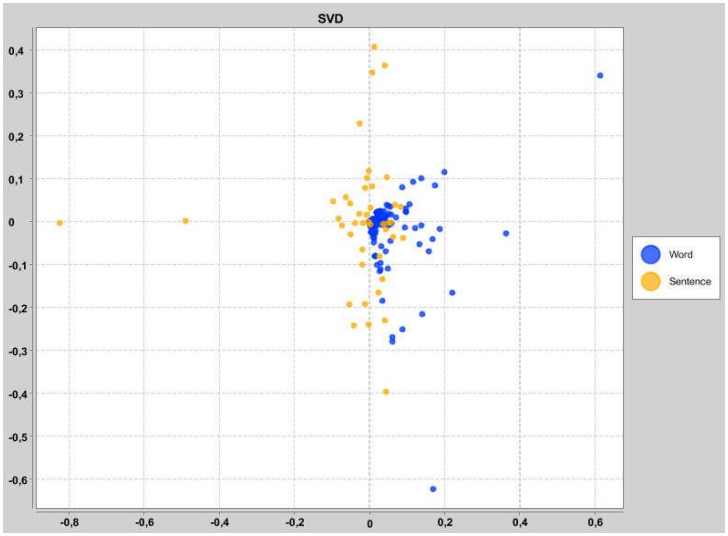

Рисунок 3 – Проекция частотной составляющей инструкции ПУЭУ-2009 и НПАОП

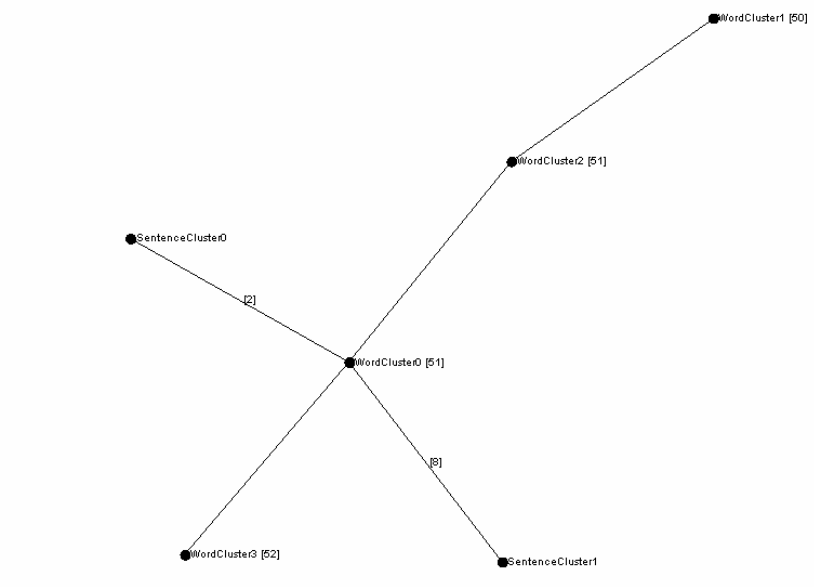

Рисунок 4 – Семантическая сеть для ПУЭУ-2009 и НПАОП

Как можно заметить, после нарушения смысловых связей между частями инструкции значительных изменений претерпела именно семантическая сеть, тогда как частотная проекция практически не изменилась. Полученные от системы семантически слабые предложения содержится в таблице 1.

Таблица 1

Семантически слабые предложения

Добавленные предложения	Исходные предложения
Необходимо содержать в порядке и постоянной готовности первичные средства пожаротушения огнетушители ящики с песком и лопатами пожарные краны и др.	Общие технические условия
	Общие требования
	Нормы качества электрической энергии в системах электроснабжения общего назначения
	В это число включаются также штепсельные розетки

Выводы и перспективы развития направления. Была сформулирована и реализована модель построения семантической сети текстовой промышленной инструкции, не требующая предварительного заполнения эталонной базы знаний и независимая от отрасли своего применения. На основе данных, полученных из семантической сети, разработан подход к определению слабых смысловых элементов и семантических ошибок в тексте инструкции. Реализована прикладная программная система на языке Java, тестирование которой демонст-

 ISSN 1562-9945

рирует возможность ее практического применения на предприятиях с целью анализа составленных инструкций и автоматизации процесса нахождения в них ошибок. Полученные в процессе тестирования результаты показывают, что система не только вернула искомое предложение, но и не добавила в результирующий набор лишних элементов – помимо тестового текста, в наборе содержаться заголовки инструкции, не связанные с конкретными предложениями, а так же предложение, требующее определенного уточнения. Помимо этого, на основе частотных портретов инструкции и семантических сетей, полученных в процессе тестирования, было установлено, что система чувствительна именно к семантическим изменениям, тогда как частотные данные практически не изменяются.

ЛИТЕРАТУРА

1. Мельчук И.А. (2012) Язык: от смысла к тексту. [Text] / Мельчук И.А.//176 с– М. 2012.

2. Кибрик А. А., И. М. Кобозева И. М. Секерина И. А. (2016). Современная американская лингвистика: Фундаментальные направления. [Text] Кибрик А. А., //М. – 2016.

3. Н.Н. Леонтьева. (2012) Автоматическое понимание текстов: системы, модели, ресурсы. [Text]/ Н.Н. Леонтьева //Москва – 2012

4. Швецов А.Н. – (2013) Система синтеза учебных тестов на основе формальных грамматик [Текст]/Швецов А.Н //журнал «Программные продукты и системы», №2(102), 2013, с 181-185.

5. Гурин Н. И. (2015) Семантическая сеть электронного учебника для диалога с виртуальным преподавателем [Текст]/Гурин Н. И., Жук Я. А. //Материалы международной научно-технической интернет конференции "Информационные технологии в образовании, науке и производстве"//БГТУ, Минск, 2015 г.

6. Тарасов Д. С. - Генерация естественного языка, парафраз и автоматическое обобщение отзывов пользователей с помощью рекуррентных нейронных сетей [Текст]//Тарасов Д. С.//«Компьютерная лингвистика и интеллектуальные технологии», №14(том 1), 2015, с 607-614//Материалы международной конференции «Диалог», 2015 г.

7. Мозговой М. В. (2012). Машинный семантический анализ русского языка и его применения [Текст]/Мозговой М. В.//СПбГУ, Санкт-Петербург –116с. – 2012г

8. V. Zakharov (2013) Russian Corpora: Comparison and Usage/ V. Zakharov//St.Petersburg State University Department of Mathematical Linguistics//The 16th International Conference TSD 2013/

9. Правила улаштування електроустановок . Міністерство енергетики та вугільної промисловості України. //Київ, 2014.

10.Нормативні акти про охорону праці (НПАОП, ДНПАОП) [Электронный ресурс]. - Access mode: https://dnaop.com/html/31549/

11.Волковский О.С., Ковылин Е.Р. Анализ современных подходов к задаче автоматической генерации текстов на естественном языке // Системные технологии. Региональный сборник межвузовских научных трудов. - 2016. №5(106) . -С. 3-12.