ПРИМЕНЕНИЕ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ ДЛЯ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ И МАРШРУТИЗАЦИИ ТЕКСТОВЫХ СООБЩЕНИЙ И ДОКУМЕНТОВ В РАБОТЕ ЕЦОР И СЛУЖБ ЭКСПЛУАТАЦИИ И РЕМОНТА

А. В. Хельвас, А. Д. Овчинников, А.А. Кузнецова, А. А. Гиля-Зетинов (Москва)

Введение

При работе сложных организационно-технических комплексов, относящихся к объектам критической инфраструктуры (далее – ОКИ) одной из важных задач является координация взаимодействия персонала, обеспечивающего поддержание эксплуатационной готовности и проведение оперативных работ на большом пространстве с использованием современных технологий поддержки принятия решений. Инструментом такого взаимодействия является Единый центр оперативного реагирования (далее – ЕЦОР). ЕЦОР обеспечивает взаимодействие нескольких дежурных диспетчерских служб (ДДС), создаваемых для управления сообщениями о жизнедеятельности и выполнении регламентов по поддержанию жизненного цикла объектов критической инфраструктуры.

Одной из основных форм взаимодействия персонала с использованием ЕЦОР является режим коротких сообщений – донесений и запросов. При этом, как правило, последовательность обработки сообщений описывается некоторыми регламентами и традиционно осуществляется с использованием специального штата диспетчеров и оперативных дежурных.

Переход к современным технологиям электронного взаимодействия заметно увеличивает количество входящих сообщений и документов для ДДС. Это повышает нагрузку на сотрудников, зоной ответственности которых является классификация входящих сообщений и их привязывание к регламентам обработки и исполнителям.

Современные технологии управления также требуют формирования набора метаданных, описывающих сообщения и документы. Простым примером такого набора метаданных может служить Дублинское ядро [1]. Использование атрибутирования сообщений, как правило, автоматически увеличивает объем работы по формированию карточек документов. При этом по мере роста объема сообщений растет и количество ошибок при их обработке.

Решить проблему оперативной обработки возрастающего объема входящих сообщений и документов может применение методов искусственного интеллекта их классификации, маршрутизации и подготовки проектов решений лиц, принимающих решения.

В качестве технологической основы для практической реализации нейросетевого подхода к анализу официальных документов выбрана платформа DeepPavlov и созданное на ее основе решение COS.AIDOC.

Решение COS.AIDOC обеспечивает на различных стадиях обработки документов выполнение следующих процедур:

– составление краткой аннотации, которая дальше может использоваться вместо полного текста при принятии решений;

– извлечение из текста документа существенных фактов (наименований организаций, адресов, фамилий должностных лиц и т.д.);

– классификация документа, определение регламента, в соответствии с которым он должен обрабатываться, подготовка проекта резолюции.

Решение задачи классификации, в свою очередь, состоит из четырех последовательно выполняемых этапов:

- предварительная обработка массива текстов;
- подбор размерности пространства признаков;
- подбор обучающей выборки и обучение классификатора;
- оценка качества результата.

Для входящих бумажных документов предварительно проводится сканирование и распознавание с помощью OCR решения ABBYY Fine Reader или Tesseract. Эта технология в настоящее время уже неплохо освоена и получила широкое распространение.

При разработке архитектуры сервиса нейросетевого анализа сообщений и документов, проходящих через ДДС ОКИ одной из проблем является то, что работа с сообщениями и документами осуществляется во многих точках – узлах графа, описывающего организационно – штатную структуру (далее – ОШС) управления ОКИ.

При этом ОШС с точки зрения маршрутов документов представляет собой не дерево, а граф с достаточно сложной структурой, образуемый как совокупность графов, описывающих маршруты документов различного типа. При этом зачастую обработка документа осуществляется не путем применения типового маршрута, а на основе решений принимаемых в каждом узле графа. Узлами графа могут служить операторы ДДС, командиры частей и подразделений МЧС, должностные лица организаций, расположенных на территории ОКИ.

Если мы ставим задачу использования нейросетевого алгоритма для маршрутизации и подготовки проекта решения по документу, то обучение сети должно осуществляться для каждого из узлов такого графа.

На рисунке 1 приведен фрагмент сети обработки документа в большой структуре с несколькими тысячами узлов обработки документов.

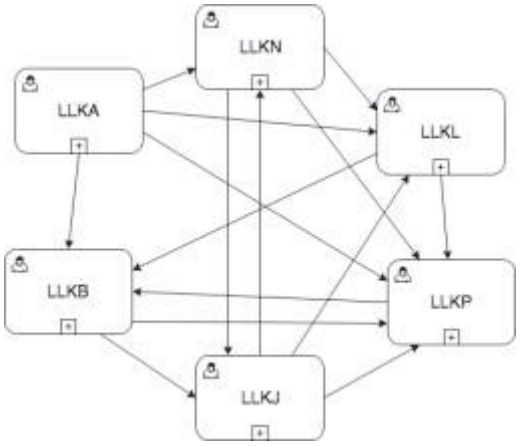

Рис. 1. Фрагмент сети обработки сообщений в ЕЦОР

Несложно заметить, что в цепочках обработки документов два разных узла могут встретиться различными способами:

LLKA ➔LLKP ➔LLKB

LLKA ➔LLKB ➔LLKJ ➔LLKP

LLKA ➔LLKB ➔ LLKP

Очевидно, что для каждого узла (например, LLKB) решение по дальнейшей маршрутизации должно основываться на собственно документе и резолюциях тех, кто работал с документом до того, как он попал к LLKB.

И, таким образом, количество обученных нейросетей для организации, граф обработки документов которой содержит N узлов должно быть равно N.

В статье предложено решение для архитектуры нейросетевого алгоритма, используемого для маршрутизации и подготовки проектов решений для обработки сообщений в ЕЦОР со сложной организационной структурой в части схем взаимодействия. При этом учтено влияние уже имеющихся результатов обработки сообщения на ранних стадиях жизненного цикла.

Также в рамках работы проведен эксперимент по обработке коротких сообщений с целью их классификации и маршрутизации на примере сообщений, поступающих в ДДС ЖКХ крупного мегаполиса. Для обучения нейросети использовалась размеченная выборка из 70 000 сообщений.

В научно – технической литературе тема применения методов машинного обучения для работы с текстовыми документами является в настоящее время одним их наиболее обсуждаемых направлений применения нейросетевых алгоритмов.

Так, например, в статьях [2,3] приведены обзоры подходов к применению нейросетевых алгоритмов для решения задачи классификации документов.

В работе [4] приведен обзор статей за 2011-2016 год по теме автоматической классификации документов, содержащий, в том числе, сравнительный анализ применения различных типов нейросетевых алгоритмов и даны рекомендации по подготовке обучающей выборки для применения CNN сетей.

Описание выбранной нами архитектуры сети приведено в статье [5]. При этом использованная технология векторизации слов (fastText) описана в статье [6].

На рисунке 2 приведено схематическое изображение стенда, на котором проводилась отладка предложенных решений.

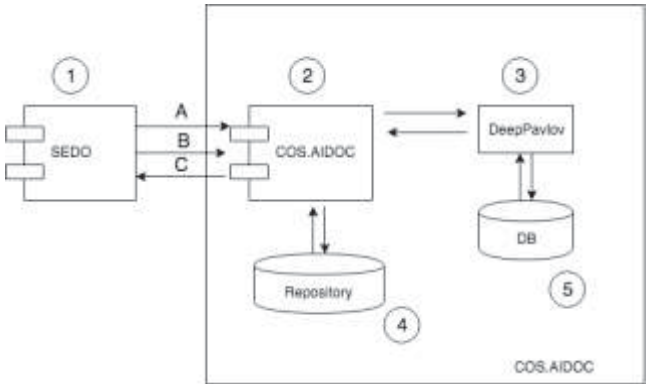

Рис. 2. Общая схема создаваемого решения

На рисунке приведены система единого центра оперативного реагирования (далее – ЕЦОР), используемая ДДС 1, интеграционный модуль COS.AIDOC 2, нейросетевая NLP платформа iPavlov 3, репозиторий результатов обучения сети для различных узлов сети 4 и хранилище NLP платформы 5.

Описание процессов информационного обмена

На рисунке 2 также приведены основные сообщения, которыми обмениваются система ЕЦОР 1 и платформа COS.AIDOC 2, здесь:

A – загрузка обучающей выборки;

B – запрос на обработку документа;

C – результат обработки запроса.

Разработаны XML схемы для структурированных электронных документов, которыми разработанный сервис обменивается с ЕЦОР.

В режиме первичного обучения осуществляется выгрузка из хранилища ЕЦОР 1 массива данных в виде архива zip формата, включающего карточки сообщений и документов в XML формате и самих сообщений и документов в текстовом формате.

Карточка документа имеет вид:

```
<?xml version="1.0"?>
<AIDOC-learn>
  <Text>XXXXXXXX</Text>
    <Classifiers>
      <Classifier class_id="Group" elementID="0.A.RQU."/>
      <Classifier class_id="Rubric" elementID="0.0.11FE.11GR.11GX."/>
      <Classifier class_id="Res_Author" elementID="0.CDMV0.CDMV5."/>
      </Classifiers>
    <Annotations>
       <Annotation Ann_typeID="1" quality="1.0">XXXXXXX</Annotation>
    </Annotations>
    <Visa>XXXXXXXXXXX</Visa>
    <Routes>
      <Route step="1" actorID="LLKA" targetID="LLKB">Текст резолюции</Route>
      <Route step="2" actorID="LLKB" targetID="LLKJ"> Текст резолюции </Route>
    </Routes>
</AIDOC-learn>
```

Архив представляет собой иерархию каталогов YYYY/MM/DD в которых находятся пары файлов XXXX.XML и XXXX.pdf (.txt) обучающей выборки.

Обучение производится для каждого узла организационной структуры. При этом в репозитории формируется набор параметров для каждого узла в каталоге, имя которого для простоты совпадает с наименованием узла в дереве.

После завершения обучения, решение переводится в режим обработки запросов.

Далее ЕЦОР направляет запрос В к системе COS.AIDOC.

```
<?xml version="1.0"?>
<AIDOC-request>
   <text url="document.pdf"> Тут содержится полный текст   </text>
   <RoutesHist>
     <Route step="1" actorID="LLKA"  targetID="LLKB">Текст резолюции</Route>
     <Route step="2" actorID="LLKB"  targetID="LLKJ"> Текст резолюции </Route>
   </RoutesHist>
</AIDOC-request>
```

Раздел <RoutesHist> содержит информацию о обработке документа на более ранних этапах обработки. Ответ на запрос передается в формате:

```
<?xml version="1.0"?>
<AIDOC-responce>
 <Classifiers>  <! - Блок
   <Classifier clas_id="1" elementID="02" probability="0.4"/>
   <Classifier clas_id="2" elementID="12" probability="1.0"/>
 </Classifiers>
 <Annotations>
     <Annotation Ann_typeID="1" quality="0.87">Annotation text 1</Annotation>
 </Annotations>
 <DataExtras>
     <Data data_typeID="1">Person 1</Data>
     <Data data_typeID="1">Person 2</Data>
     <Data data_typeID="2">Journal 1</Data>
 </DataExtras>
 <Visa>Resolution text </Visa>
 <Routes>
     <Route step="1" actorID="ivanov" probability="0.78">Resolution 1 </Route>
     <Route step="1" actorID="petrov" probability="0.22">Resolution 2 </Route>
 </Routes>
</AIDOC-responce>
```

Ответ содержит следующие блоки:

<Classifiers> – результаты классификации предложенного документа;

<Annotations> – варианты аннотации документа с оценкой предполагаемого качества quality;

<DataExtras> – извлеченные из текста данные (при этом атрибут data_typeID описывает тип извлекаемых из текста данных);

<Resolution> – проект резолюции;

<Executors> – перечень предполагаемых исполнителей по документу (при этом атрибут main описывает, является ли исполнитель ответственным, а атрибут probability – оценку правильности выбора исполнителя системой).

Экспериментальная оценка показателей

Разработанное решение было апробировано на основе массива размеченных данных портала ЖКХ крупного мегаполиса.

Размер обучающей выборки 70000 документов.

Полученные значения показателей приведены в таблице.

Таблица 1. Показатели работы решения

	Наименование показателя	Значение	Комментарий
1	Время обработки документа, с	15.2	Без учета времени работы OCR
2	Вероятность правильного определения типа документа	0.76	По данным выборки ЕЦОР ЖКХ из 70 000 записей
3	Вероятность правильной маршрутизации	0.87	По данным выборки ЕЦОР ЖКХ из 70 000 записей

Заключение

Таким образом, нами разработана архитектура решения, обеспечивающего маршрутизацию документов в организациях, описываемых сложным графом обработки документов.

Предложено при проведении обучения использовать карточки документа, форматированные для определения решения, принимаемого в конкретном узле графа, описывающего организационную структуру.

Разработаны структуры данных для сообщений, которыми обменивается система электронного документооборота и нейросетевая система маршрутизации и классификации документов.

Проведены эксперименты и подтверждена высокая эффективность предложенного подхода.

Литература

1. «Dublin Core Metadata Initiative» official site, http://dublincore.org.
2. **Епрев А.С.** Автоматическая классификация текстовых документов // Математические структуры и моделирование. 2010. Вып. 21. С. 65-81.
3. **Батура Т.В.** Методы автоматической классификации текстов // Программные продукты и системы. 2017. Том. 1(30). Стр. 85–99.

4. **Белоус Р.О., Чернятина Ю.А.** Применение нейронных сетей в задачах обработки текстовых данных // Научно-технический вестник информационных технологий, механики и оптики. 2008.
5. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, page 1746-1751. (2014).
6. **Bojanowski P., Grave E., Joulin A., Mikolov T.** (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135-146.