М. В. Наталевич,
студент III курса Института бизнеса БГУ
Научный руководитель:
кандидат экономических наук, доцент
Т. В. Прохорова

РЕШЕНИЕ ЗАДАЧИ НОРМАЛИЗАЦИИ ТЕКСТА ДЛЯ ЭЛЕКТРОННОГО ВЕДЕНИЯ БИЗНЕСА

Стремительный рост деловой активности в обществе, постоянная интенсификация потоков информации, внедрение инновационных технологий в различных сферах деятельности человека, приводит к значительному увеличению объема документов. Современные технологии позволяют формировать, распространять и копировать документы значительно быстрее по сравнению с предшествующими десятилетиями.

Сегодня в республике активно развиваются новые инструменты для эффективного обеспечения управленческих процессов. В частности, внедряется использование электронного документа, электронно-цифровой подписи, системы электронного документооборота. В декрете «О развитии цифровой экономики» 2017 г. [1], а также постановлении Совета Министров Республики Беларусь «Об утверждении Государственной программы развития цифровой экономики и информационного общества» 2016 г. [2] определены векторы создания и поддержки необходимых условий для развития в стране цифровой экономики.

Актуальность работы определяется необходимостью создания программного обеспечения, предназначенного для обработки документов и электронного ведения бизнеса.

Современные достижения в области компьютерных и информационных технологий обусловили возможность создания в различных областях человеческой деятельности автоматизированных систем обработки информации. Целью данного исследования является решение задачи нормализации текста для электронного ведения бизнеса.

Бизнес-документы содержат большое число элементов текста, требующих нормализации. Например, при указании суммы в текстах договоров в скобках приводится расшифровка в нормализованном виде. Алгоритм, обученный для решения задачи, сможет сгенерировать такую расшифровку самостоятельно. Более того, такая система сможет исследовать более старые документы и найти ошибки, допущенные при их ручном наборе. В условиях электронного документооборота требуется гораздо меньше затрат на перестройку документооборота при изменении внешних условий, например, требований по изменению формы отчетности.

Наша задача по нормализации текста на данном этапе работы заключается в переводе письменного текста в произносимую форму. В тексте могут встречаться необычные элементы, например, «3 кг» или «1200 р.». Данные выражения после нормализации будут превращены в «три килограмма» и «тысяча двести рублей», соответственно. Помимо единиц измерения и валют, задача нормализации распознает более 10 видов необычных данных, таких как сокращения, числительные, пунктуационные знаки, даты, время и др.

Во многих случаях нормализация может быть проведена единственным образом (например, «ул.» может означать только слово «улица»). Но в некоторых случаях разрешение сокращений неоднозначно (например, сокращение «г.» может означать как «город», так и «год»). В таких ситуациях алгоритм должен обработать контекст, чтобы понять, какая нормализация является корректной.

Для решения задачи программе предоставляется большой корпус текстов, для которых известна нормализованная форма. По этим данным алгоритм изучает закономерности и правила преобразования элементов текста в произносимую форму.

Задачу нормализации текстов принято рассматривать в контексте задачи синтеза речи. Первые работы по синтезу речи датируются 1987 г., и с того времени данное направление активно развивается. В связи с появлением алгоритмов машинного обучения задача нормализации, как и другие задачи, связанные с обработкой текстов, стали развиваться особенно активно [3].

Эффективное решение задачи нормализации текстов и программы, основанные на таком решении, помогут уменьшить затраты времени на ведение документации в компаниях Беларуси. Более того, использование систем на основе решения задачи нормализации может избавить от потенциальных ошибок, совершенных по вине человека. Такой подход согласуется с вектором политики страны в цифровой экономике и современными требованиями к ведению бизнеса.

Список использованных источников

1. О развитии цифровой экономики [Электронный ресурс] : Декрет Президента Респ. Беларусь, 21 декабря 2017 г. № 8 // Нац. правовой интернет-портал Респ. Беларусь. – Режим доступа: http://president.gov.by/ru/official_documents_ru/view/dekret-8-ot-21-dekabrja-2017-g-17716/. – Дата доступа: 18.04.2019.

2. Об утверждении Государственной программы развития цифровой экономики и информационного общества на 2016–2020 годы [Электронный ресурс] : постановление Совета Министров Респ. Беларусь 23 марта 2016 г., № 235 // Нац. правовой интернет-портал Респ. Беларусь. – Режим доступа: http://pravo.by/document/?guid=3871&p0=C21600235. – Дата доступа: 18.04.2019.

3. RNN Approaches to Text Normalization: A Challenge [Electronic resource]. – Mode of access: https://arxiv.org/abs/1611.00068. – Date of access: 18.04.2019.