Методы выявления связей между нормативно-правовыми документами*

Д. А. Девяткин[I], А. Т. Софронова[II], И. В. Соченков[I]

[I] Федеральный исследовательский центр «Информатика и управление» РАН, Москва, Россия
[II] ООО «Технологии системного анализа», Москва, Россия

Аннотация. В статье предложен новый метод выявления связей (явных и неявных) между нормативно-правовыми документами, а также проведена его экспериментальная оценка на корпусе юридических текстов, посвященных регулированию в сфере информационных технологий. В основе метода лежат алгоритмы поиска тематически похожих документов и оценки семантического сходства предложений, позволяющие оценивать близость текстов с учетом не только лексики, но также выявленных синтаксических связей и семантических отношений.

Ключевые слова: извлечение информации из нормативно-правовых документов, выявление неявных связей, реляционно-ситуационный анализ.

DOI 10.14357/20718594190407

Введение

Участники нормотворческой деятельности порождают и анализируют огромный объем информации (например, в Апелляционную коллегию по гражданским делам Московского городского суда поступает более 50 тыс. дел в год [1]). В то же время работа с этой информацией автоматизирована слабо. С цифровыми системами работают не все субъекты законотворчества, а функции этих систем сводятся исключительно к систематизации законодательства в виде электронных справочников, таких как «Кодекс» или «КонсультантПлюс». При этом перспективные методы извлечения информации практически не применяются [2, 3].

Особенностью нормативно-правовых документов является значительное количество явных и неявных связей: ссылок, а также дословно совпадающих и перефразированных текстовых фрагментов из других документов. Автоматизация выявления этих связей может изменить подход к разработке и принятию нормативных правовых актов, сделать более удобным их практическое применение и мониторинг. Кроме того, решение этой задачи позволит эффективнее систематизировать базы нормативных документов.

В работе предложен новый метод автоматического выявления связей (явных и неявных) между нормативно-правовыми документами, а также проведена его экспериментальная оценка на корпусе из 1,6 тыс. юридических документов, посвященных правовому регулированию в сфере информационных технологий. В основе метода лежат алгоритмы поиска тематически похожих документов и оценки семантического сходства предложений [4,5], позволяющие оценивать близость текстов с учетом не только лексики, но также выявленных синтаксических связей и семантических отношений.

* Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 18-29-16022 «мк».

✉ *Девяткин Дмитрий Алексеевич. E-mail: devyatkin@isa.ru*

1. Методы интеллектуального анализа нормативно-правовых документов

Работы в области методов интеллектуального анализа нормативно-правовых документов посвящены решению следующих задач: информационный поиск юридических документов, извлечение информации (сущностей и связей), выявление связей между документами, диалоговые интеллектуальные агенты-помощники для консультирования пользователей при решении юридических вопросов.

Среди исследований, посвященных извлечению информации из юридических текстов, необходимо упомянуть статью [6], в которой предложен метод автоматического выявления изменений и дополнений в законодательных актах и нормативных документах. Под изменениями и дополнениями понимается описание того, как был модифицирован какой-либо документ. Результатом работы метода является консолидированный текст — это версия нормативного текста, содержащая все изменения. Разметка изменений и дополнений подразумевает:

1) определение типа изменения (например, удаление или замена) изменяемого текста (например, другой закон, указ и т.д.), часть такого документа, на которую повлияло изменение (структурная часть, например, «Статья 56» или фрагмент текста, например, словосочетание «пять лет»);

2) создание набора метаданных, который компактно описывает модификацию. Чтобы автоматически выполнить эти два шага, применяются методы обработки естественного языка. В статье описывается система, предназначенная для автоматической разметки итальянских юридических текстов, содержащих изменения и дополнения. Система опирается на синтаксический и поверхностный (shallow) семантический анализ юридических текстов на естественном языке. Извлечение изменений и дополнений проводится в три этапа. Сначала извлекаются целевые фрагменты текста документа, после чего, с помощью библиотеки TUP [7], проводится синтаксический анализ фрагментов. На третьем шаге выполняется семантический анализ, результаты которого используются для определения ролей и связей между выделенными фрагментами.

В исследовании [8], представлена семантическая модель правовых ресурсов, включающая в себя аннотации и обоснования нормативных положений, а также подход к ее построению. В этой работе делается предположение, что логические отношения между положениями нормативно-правовых документов могут быть заданы как аксиомы о типах и атрибутах положений, и как фундаментальные отношения «право/обязанность», «свобода/отсутствие свободы», «власть/ответственность». Для описания таких отношений используется язык OWL-DL [9]. Недостаток этого подхода состоит в необходимости использовать размеченный юридический корпус, а также классифицировать его тексты, опираясь исключительно на семантику содержащихся в них положений. Эта задача зачастую не может быть решена автоматически и, следовательно, необходимо привлечение экспертов. Близкий подход к решению этой задачи изложен в статье [10]. Особенностью этого подхода является совместное использование полного синтаксического и поверхностного семантического анализа естественного языка, что позволяет дополнить извлекаемые сущности информацией о семантических отношениях между ними.

Пример практического использования семантических моделей в области юриспруденции приведен в работе [11], в которой представлена система управления юридическими документами и знаниями Eunomos. Система позволяет автоматизировать решение пяти основных проблем современного процесса нормотворчества.

1) Законы четко не классифицированы по сферам применения.

2) Множественные юрисдикции (ограниченная применимость того или иного документа).

3) Большой объем текстов и большое количество документов.

4) Ограниченная доступность юридических документов.

5) Тексты нормативно-правовых документов могут периодически обновляться.

Проблема увеличения объема и количества документов решается в этой системе путем создания большой базы данных нормативных актов, преобразованных в XML и автоматически загружаемых с законодательных порталов, которые регулярно комментируются и обновляются. Проблема выявления сферы применения тех или иных законодательных актов решается с помощью методов тематической классифика-

ции текстов статей, позволяющей пользователям просматривать только те разделы законодательства, которые относятся к их сфере интересов. Проблема многозначности юридических терминов решается с помощью многоуровневых обновляемых онтологий, в которых концепты соответствуют различным понятиям и определениям.

В статье [12] рассматривается задача перехода от юридического текста на естественном языке к соответствующему набору машиночитаемых правил. Авторами предлагается предварительный подход, который объединяет различные методы обработки естественного языка для извлечения правил из юридических документов. В рамках этого подхода решаются задачи извлечения сущностей и связей из юридических текстов. Для извлечения сущностей совместно используется лингвистическая информация, содержащаяся в онтологии WordNet [13], и сформированные вручную правила, учитывающие информацию о синтаксических связях. Для извлечения зависимостей между сущностями используются логические правила. Такой комбинированный подход позволяет эффективно извлекать машиночитаемые правила из юридических документов. Для извлечения синтаксических связей используется библиотека Stanford NLP [14], а для формирования множества логических правил – библиотека Prover9 [15]. Близкий подход используется и в [16]: извлечение именованных сущностей и связей между ними осуществляется с помощью sketch engine [17], а затем выполняется автоматизированная фильтрация результатов и построение онтологии в формате RDF.

В статье [18] предложен новый метод извлечения именованных сущностей и отношений между ними из нормативных документов – вместо использования методов сопоставления с образцом, основанных на лексико-синтаксических шаблонах, применяется машинное обучение, причем в качестве признаков используются синтаксические зависимости между терминами. В качестве метода машинного обучения используется SVM, который обучается на размеченном вручную законодательном корпусе.

В работе [19] предложен подход к классификации юридических документов, в котором документ, относящийся к нескольким классам, рассматривается как набор из нескольких документов, каждый из которых связан только с одним классом. Для извлечения признаков документов, используемых в ходе классификации, использован подход к оценке весов терминов, представленный в [20] и лингвистический анализатор TULE [21]. Для оценки качества классификации использовался корпус JRC-Acquis, содержащий более 20 тыс. документов, каждый из которых в среднем помечен шестью категориями. В результате, предложенный метод позволил добиться более высокого качества, чем стандартные подходы к многозначной (multilabel) классификации.

В области создания автоматических диалоговых агентов для проведения юридических консультаций, необходимо отметить метод, представленный в работе [22]. В основе него лежит многослойная нейронная сеть с архитектурой «кодировщик-декодировщик», внутренним рекуррентным слоем (Long-Short Term Memory LSTM [23]) с механизмом внимания. Для обучения использовались 1200 пар вопросов и ответов, извлеченных из диалогов реальных юридических консультантов со своими клиентами. Экспертный анализ работы агента подтвердил, что качество его консультирования превосходит результаты, полученные с помощью агентов, основанных на правилах.

Значительный массив работ посвящен исследованиям в области поиска юридической информации. Например, в статье [24] выделяются следующие подходы к оценке релевантности при поиске нормативных документов.

1. Алгоритмическая релевантность (наличие в ответе запрашиваемой информации).
2. Тематическая релевантность (соответствие тематики запроса и ответа).
3. Ситуационная релевантность (соответствие описываемой в запросе юридической ситуации и найденных документов).
4. Полезность для принятия решений.
5. Юридическая значимость результатов.

Очевидно, что перспективные методы поиска юридической информации должны учитывать перечисленные типы релевантности.

Еще одно методологическое исследование посвящено постановке задачи поиска юридических текстов, в ней предлагаются методы оценки и специализированные поисковые стратегии [25]. Экспериментальная оценка предложенных стратегий проводилась на своде решений Верховного суда США. Наилучшие оценки точно-

сти и полноты получила стратегия, предполагающая автоматическую настройку глубины поиска с помощью регрессионных моделей.

Непосредственно методам информационного поиска юридических текстов посвящена статья [26], в которой предлагается совместно применять многослойные нейронные сети, предварительно обученные без учителя, BERT [27] (языковые модели) и ранжирование с помощью меры BM25. В итоге такой комбинированный подход показал лучшую точность поиска, чем BERT или BM25 по отдельности.

Задача выявления связей между нормативно-правовыми документами рассматривается в работе [28]. В ней для предсказания ссылок на юридические документы используется обучение с подкреплением. В ходе обучения подкрепляются случаи корректного предсказания ссылок на другие документы. В статье [29] для выявления связей юридических документов используется языковая модель BERT, позволяющая неявно учитывать разнородные лингвистические признаки и контекст анализируемых фрагментов текста. Экспериментальная оценка такого подхода на размеченном корпусе показала его преимущество перед методами, основанными на применении правил.

По итогам обзора, необходимо отметить, что для автоматического анализа нормативно-правовых документов и выявления связей между ними используются либо методы, явно учитывающие разнородные признаки, извлеченные с помощью полного лингвистического анализа, в том числе, синтаксические и семантические связи, либо предобученные языковые модели.

2. Метод выявления связей между юридическими документами

Разработанный метод выявления связей между юридическими документами состоит из следующих шагов (Рис. 1).

1. Полный лингвистический анализ текста, включающий токенизацию, морфологический анализ, построение синтаксических деревьев зависимостей, реляционно - ситуационный анализ [30]. Для выполнения этих операций используется анализатор, разработанный в ФИЦ ИУ РАН [31].

2. Выявление явных ссылок на документы с использованием множества правил, составленных с применением языка регулярных выражений. Выделение явных ссылок проводилось в два этапа, на первом выявлены предложения, которые были оформлены в соответствии с правилами оформления ссылок, так, к примеру, выделялись предложения, имеющие в своем составе следующие словосочетания: «Согласно статье (ям)» или «В соответствии с». На втором этапе из найденных предложений были выделены названия законов. Далее по базе метаданных выполняется поиск документов, на которые указывают выявленные ссылки.

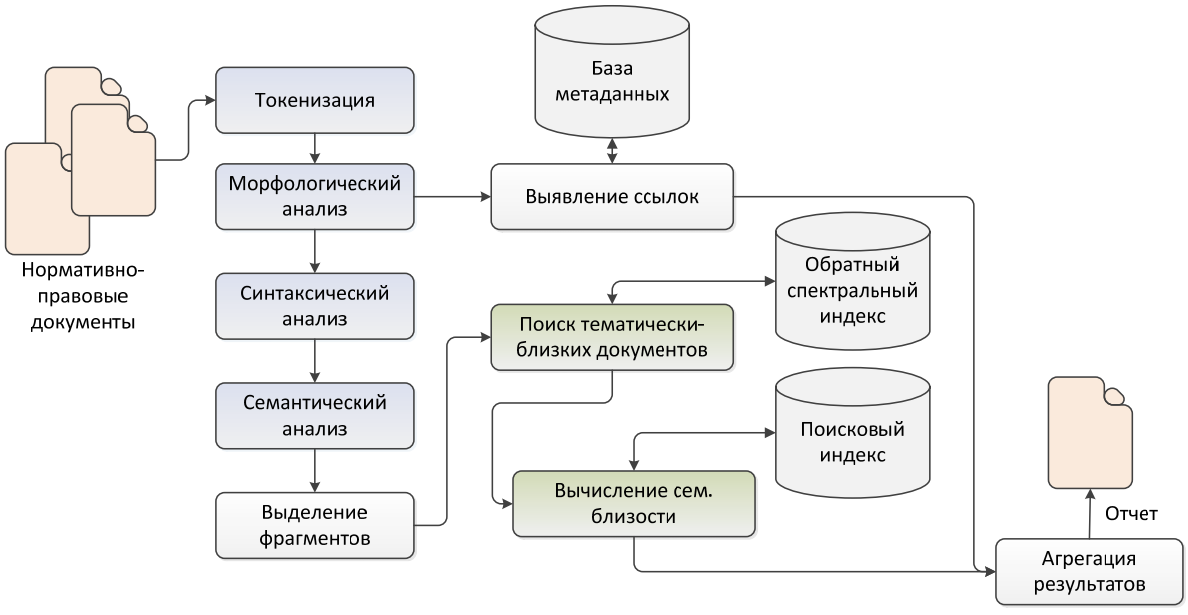

Рис. 1. Схема выявления неявных связей между юридическими документами

3. Разбиение анализируемого текста d на фрагменты $d = \{\tau\}$. В случае, если текст имеет явно заданную структуру в виде параграфов или пунктов, то каждый отдельный элемент этой структуры может рассматриваться в качестве фрагмента. Иначе текст разбивается на пересекающиеся фрагменты фиксированного размера.

4. Выявление неявных связей. Для решения этой задачи используется адаптированный метод выявления дословных и перефразированных текстовых заимствований, предложенный ранее в [32]. С помощью этого метода выполняется поиск документов, тематически близких к фрагментам анализируемого текста. Используется мера тематической близости, предложенная в [4]. Благодаря несимметричности этой меры, размер и тематическая неоднородность документов, по которым производится поиск, не влияют на результат. Для вычисления оценки близости будем использовать косинусную оценку тематического сходства SIM_{cos} или оценку тематической схожести по Хэммингу SIM_{Ham} (опционально):

$$SIM_{cos}(\tau, d_j) = \\ = \frac{\sum_{w \in \widehat{W}(\tau) \cap \widehat{W}(d_j)} v(w,\tau,c) v(w,d_j,c)}{\sqrt{\sum_{w \in \widehat{W}(\tau) \cap \widehat{W}(d_j)} \left(v(w,\tau,c)\right)^2} \sqrt{\sum_{w \in \widehat{W}(\tau) \cap \widehat{W}(d_j)} \left(v(w,d_j,c)\right)^2}}, \quad (1)$$

$$SIM_{Ham}(\tau, d_j) = \\ = 1 - \frac{\sum_{w \in \widehat{W}(\tau) \cap \widehat{W}(d_j)} |v(w,\tau,c) - v(w,d_j,c)|}{\sum_{w \in \widehat{W}(\tau) \cap \widehat{W}(d_j)} \left(v(w,\tau,c) + v(w,d_j,c)\right)}, \quad (2)$$

где τ – фрагмент-эталон; d_j – документ, сходство которого с эталоном оценивается; $c = \{d\}$ – коллекция юридических текстов; $\widehat{W}(\tau)$, $\widehat{W}(d_j)$ – множества лексических единиц фрагмента τ и документа d_j, соответственно; $v(w,d,c)$ – числовая функция, задающая вес лексической единицы w в тексте (фрагменте текста) d из коллекции c.

5. Вычисление оценки близости, предложенной в [5]. Оценивается сходство между фрагментами анализируемого текста и найденных документов с учетом синтаксических связей и семантических отношений. Общая оценка сходства фрагмента τ исходного текста и фрагмента ε сопоставляемого текста определяется взвешенной суммой величин:

$$Sim(\tau, \varepsilon) = \sum_{n=1}^{5} \lambda_n I_n(\tau, \varepsilon), \quad (3)$$

где

$$I_1(\tau, \varepsilon) = \frac{\sum_{w \in \widehat{W}(\tau) \cap \widehat{W}(\varepsilon)} |SemR_w^{\varepsilon} \cap SemR_w^{\tau}|}{|SemRoles^{\tau}|}, \quad (4)$$

$SemR_w^{\tau} = \{a \in Roles | \quad \exists w' \in \widehat{W}(\tau), \exists x \in R: \quad < w, w' > \in \Omega_x^{\tau} \,\&\, < w', a > \in SemRoles^{\tau}\}$ – множество значений синтаксем, приписанных во фрагменте τ тем словоупотреблениям w', которые связаны в этом тексте со словоупотреблением w всевозможными семантическими связями.

Бинарное отношение $SemRoles^{\tau} \subseteq \widehat{W}(\tau) \times Roles$ сопоставляет словоупотреблениям текста семантические значения соответствующих синтаксем из конечного множества значений $Roles$.

Бинарное отношение $\Omega^{\tau} \subseteq \widehat{W}(\tau) \times \widehat{W}(\tau)$ задает семантически связанные словоупотребления в тексте.

$$I_2(\tau, \varepsilon) = \frac{\sum_{<w,w'> \in N_{Syn}(\tau,\varepsilon)} v(w,\tau,c)}{\sum_{w \in \{w \in \widehat{W}(\tau) | \exists w' \in \widehat{W}(\tau): <w,w'> \in \Sigma^{\varepsilon}\}} v(w,\tau,c)}. \quad (5)$$

Множество $N_{Syn}(\tau, \varepsilon) = \{< w^{\tau}, w^{\varepsilon} > \in (\widehat{W}(\tau) \cap \widehat{W}(\varepsilon))^2 | \exists \tilde{w}^{\varepsilon}, \tilde{w}^{\tau} \in \widehat{W}(\tau) \cap \widehat{W}(\varepsilon), \exists z \in SR: < \tilde{w}^{\varepsilon}, w^{\varepsilon} > \in \Sigma_z^{\varepsilon} \,\&\, < \tilde{w}^{\tau}, w^{\tau} > \in \Sigma_z^{\tau}\}$ содержит пары словоупотреблений в эталонном фрагменте τ и сопоставляемом фрагменте ε, для которых совпадают (по нормальным формам) главные и зависимые слова, а сами словоупотребления связаны в контексте исходного и сопоставляемого фрагментов однотипными синтаксическими связями:

$$< w^{\varepsilon}, \tilde{w}^{\varepsilon} > \in \Sigma_z^{\varepsilon} \;\&\; < w^{\tau}, \tilde{w}^{\tau} > \in \Sigma_z^{\tau};$$

$$I_3(\tau, \varepsilon) = \frac{|\rho(\tau,\varepsilon)|}{|SemRoles^{\tau}|}. \quad (6)$$

Множество $\rho(\tau, \varepsilon) = \{< w^{\varepsilon}, a > \in SemRoles^{\varepsilon} | w^{\varepsilon} \in \widehat{W}(\varepsilon) \& a \in Roles \& \exists w^{\tau} \in \widehat{W}(\tau) < w^{\varepsilon}, w^{\tau} > \in (\widehat{W}(\tau) \cap \widehat{W}(\varepsilon))^2, < w^{\tau}, a > \in SemRoles^{\tau}\}$, содержит словоупотребления во фрагменте эталона и во фрагменте сопоставляемого текста, у которых совпадают семантические значения.

В качестве I_4 может использоваться SIM_{cos} или SIM_{Ham}, в качестве дополнительного критерия связности текстов I_5 – доля общих концептов юридических онтологий, таких как LKIF [33] или ее адаптаций для РФ [34].

Множество $L = \{\lambda_i | i = 1 \ldots 5\}$ задает набор параметров метода, определяющих вклад каждого из критериев оценки близости в итоговую величину. В случае если вычисленная оценка

близости с некоторым документом превышает заранее заданный порог, фрагмент считается неявно связанным с этим документом.

6. Агрегация выявленных связей (явных и неявных) и формирование итогового отчета.

3. Оценка метода выявления связей между нормативно-правовыми документами

Экспериментальная оценка предложенного метода проводилась на размеченном корпусе, включающем в себя более 1,6 тыс. юридических документов, посвященных правовому регулированию в сфере информационных технологий. Результаты оценки приведены в Табл. 1.

Из таблицы видно, что предложенный метод обладает относительно высокой полнотой ($R \geq 0.9$). При этом детальный анализ результатов тестирования показывает, что к снижению показателя точности (P) приводит наличие в юридических текстах фрагментов, незначимых с точки зрения тематики документа, но играющих большую роль при связывании. К таким фрагментам относятся, например, регистрационные номера и даты. Поэтому, в дальнейшем, планируется использовать гибридный показатель связности нормативно-правовых документов, учитывающий сходство подобных фрагментов. В качестве основного варианта применения метода представляется его использование в составе автоматизированных систем

Табл. 1. Результаты экспериментальной оценки метода выявления связей

Оценки	Явные ссылки	Неявные ссылки
Precision	0,78	0,79
Recall	0,90	0,98
F_1	0,84	0,87

поддержки юридической деятельности, в которых конечное решение о наличии связи между документами принимает оператор. Вариант представления результатов работы метода в подобной системе представлен на Рис. 2.

Заключение

Предложенный метод учитывает разнородные лингвистические признаки и позволяет достаточно точно находить явные и неявные связи между нормативно-правовыми документами. В перспективе на основе этого метода возможно создание систем поддержки юридической деятельности, которые позволят существенно сузить круг оцениваемых актов, сократить временной интервал оценки предлагаемой информации, а также повысить качество нормотворческого процесса в целом. Помимо применения для решения прикладных задач, этот метод может быть использован в качестве базового при экспериментальной оценке других подходов к решению задачи выявления связей.

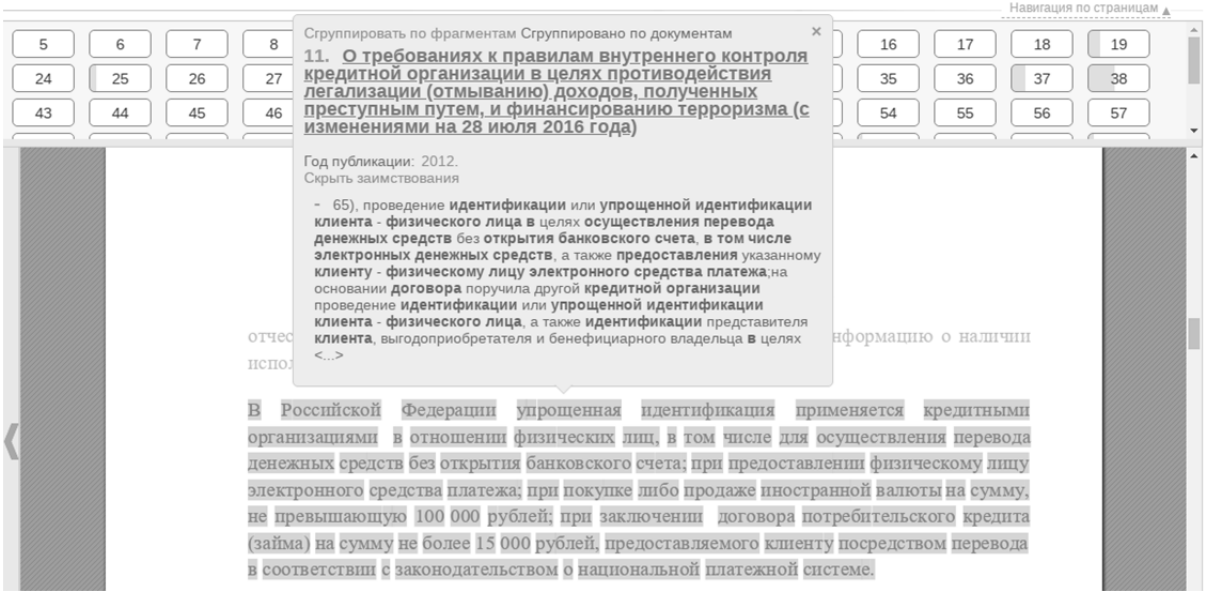

Рис. 2. Отображение выявленных неявных связей

Литература

1. Сравнительные статистические данные работы Московского городского суда за 2017 г., размещенные на официальном сайте, электронный ресурс: https://www.mos-gorsud.ru/getGalleryImage/8083cedf-6812-4bc4-b597-4f5747ee2791
2. Неживых И.А., Автоматизация законотворчества: новая профессия СЭД, 2016, эл. ресурс: http://www.iksmedia.ru/articles/5290923-Avtomatizaciya-zakonotvorchestva.html#ixzz5F0p6A6h
3. Казиев В.М. Введение в правовую информатику. М.: НОИ Интуит. 2016.
4. Суворов Р. Е., Соченков И. В. Определение связанности научно-технических документов на основе характеристики тематической значимости //Искусственный интеллект и принятие решений. 2013. № 1. С. 33-40.
5. Zubarev D., Sochenkov I. Using Sentence Similarity Measure for Plagiarism Source Retrieval //CLEF (Working Notes). 2014. P. 1027-1034.
6. Lesmo L., Mazzei A., Radicioni D. Extracting Semantic Annotations from Legal Texts. 2009.
7. Lesmo L. Use of semantic information in a syntactic dependency parser //International Workshop on Evaluation of Natural Language and Speech Tool for Italian. – Springer, Berlin, Heidelberg. 2012. P. 13-20.
8. Semantic Model for Legal Resources: Annotation and Reasoning over Normative Provisions; Enrico Francesconi. 2016.
9. McGuinness D. L. et al. OWL web ontology language overview //W3C recommendation. Т. 10. № 10. P. 2004.
10. Combining NLP Approaches for Rule Extraction from Legal Documents; Mauro Dragoni, Serena Villata, Williams Rizzi, Guido Governatori. 2017.
11. Eunomos, a legal document and knowledge management system for the Web to provide relevant, reliable and up-to-date information on the law; Guido Boella, Luigi Di Caro, Llio Humphreys, Livio Robaldo, Piercarlo RossiLeendert van der Torre. 2016.
12. TULSI: an NLP system for extracting legal modificatory provisions; Leonardo Lesmo, Alessandro Mazzei, Monica Palmirani, Daniele P. Radicioni. 2012.
13. Miller G. A. WordNet: a lexical database for English //Communications of the ACM. 1995. Т. 38. № 11. P. 39-41.
14. Manning C. et al. The Stanford CoreNLP natural language processing toolkit //Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. 2014. P. 55-60.
15. McCune W. Release of prover9 //Mile High Conference on Quasigroups, Loops and Nonassociative Systems, Denver, Colorado. 2005.
16. Martín Chozas P. Towards a Linked Open Data Cloud of language resources in the legal domain: дис. – ETSI_Informatica. 2018.
17. Kilgarriff A. et al. The Sketch Engine: ten years on //Lexicography. 2014. Т. 1. № 1. P. 7-36.
18. Boella G., Di Caro L., Robaldo L. Semantic relation extraction from legislative text using generalized syntactic dependencies and support vector machines //International Workshop on Rules and Rule Markup Languages for the Semantic Web. – Springer, Berlin, Heidelberg. 2013. P. 218-225.
19. Boella G. et al. Linking legal open data: breaking the accessibility and language barrier in european legislation and case law //Proceedings of the 15th International Conference on Artificial Intelligence and Law. ACM. 2015. P. 171-175.
20. Salton G., Wong A., Yang C. S. A vector space model for automatic indexing. Commun. ACM, 18. November 1975. P.613–620.
21. Lesmo L. The Turin University Parser at Evalita 2009. Proceedings of EVALITA, 9.2009.
22. John A. K. et al. Legalbot: a deep learning-based conversational agent in the legal domain //International Conference on Applications of Natural Language to Information Systems. Springer, Cham. 2017. P. 267-273.
23. Hochreiter S., Schmidhuber J. Long short-term memory //Neural computation. 1997. Т. 9. № 8. P. 1735-1780.
24. Van Opijnen M., Santos C. On the concept of relevance in legal information retrieval //Artificial Intelligence and Law. 2017. Т. 25. № 1. P. 65-87.
25. Livermore M. A. et al. Law Search as Prediction //Virginia Public Law and Legal Theory Research Paper. 2018. № P. 2018-61.
26. Gain B. et al. IITP in COLIEE@ ICAIL 2019: Legal Information Retrieval using BM25 and BERT. 2019.
27. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. 2018.
28. Sun B. Information Structure Parsing for Chinese Legal Texts: A Discourse Analysis Perspective //International Journal of Technology and Human Interaction (IJTHI). 2019. Т. 15. № 1. P. 46-64.
29. Shaffer R., Mayhew S. Legal Linking: Citation Resolution and Suggestion in Constitutional Law //Proceedings of the Natural Legal Language Processing Workshop 2019. 2019. P. 39-44.
30. Осипов Г. С., Смирнов И. В., Тихомиров И. А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения //Искусственный интеллект и принятие решений. 2008. Т. 2.
31. Daniil Larionov, Artem Shelmanov, Elena Chistova and Ivan Smirnov. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates // Proceedings of Recent Advances of Natural Language Processing. 2019. P. 620-630.
32. Sochenkov I. et al. Exactus like: Plagiarism detection in scientific texts //European conference on information retrieval. Springer, Cham. 2016. P. 837-840.
33. Hoekstra R. et al. The LKIF Core Ontology of Basic Legal Concepts //LOAIT. 2007. Т. 321. P. 43-63.
34. Воронина И. Е., Пигалкова Е. А. Создание базовой онтологии для российской системы права на основе онтологии LKIF-Core //Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2010. № 1. С. 154-159.

Methods for Identifying Links Between Regulatory Documents

D. A. Devyatkin[I], A. T. Sofronova[II], I. V. Sochenkov[I]

[I] Federal Research Center «Computer Science and Control» of Russian Academy of Sciences, Moscow, Russia
[II] LLC «Technologies for Systems Analysis», Moscow, Russia

Abstract. The article proposes a new method for relationship detection (explicit and implicit) between legal documents, as well as its experimental assessment on the corpus of legal texts on information technology regulation. The method considers deep linguistic features and allows estimating topic, syntax and semantic similarity of legal texts.

Keywords: information extraction from legal documents, implicit relationship detection, relational-situational analysis.

DOI 10.14357/20718594190407

References

1. Comparative statistics of the work of the Moscow City Court for 2017, Available at: https://www.mos-gorsud.ru/getGalleryImage/8083cedf-6812-4bc4-b597-4f5747ee2791
2. Nezhivykh I.A., Legislation Automation: A New Profession in the SED, 2016, Available at: http://www.iksmedia.ru/articles/5290923-Avtomatizaciya-zakonotvorchestva.html#ixzz5F0p6A6h.
3. Kaziev V.M., Introduction to Legal Informatics, M.: NOI Intuit, 2016.
4. Suvorov R. E., Sochenkov I. V. Determination of the connectedness of scientific and technical documents based on the characteristics of thematic significance // Artificial Intelligence and Decision Making. – 2013. – № 1. – p. 33-40.
5. Zubarev D., Sochenkov I. Using Sentence Similarity Measure for Plagiarism Source Retrieval //CLEF (Working Notes). – 2014. – C. 1027-1034.
6. Lesmo L., Mazzei A., Radicioni D. Extracting Semantic Annotations from Legal Texts. – 2009.
7. Lesmo L. Use of semantic information in a syntactic dependency parser //International Workshop on Evaluation of Natural Language and Speech Tool for Italian. – Springer, Berlin, Heidelberg, 2012. – C. 13-20.
8. Semantic Model for Legal Resources: Annotation and Reasoning over Normative Provisions; Enrico Francesconi; 2016.
9. McGuinness D. L. et al. OWL web ontology language overview //W3C recommendation. – 2004. – T. 10. – № 10. – C. 2004.
10. Combining NLP Approaches for Rule Extraction from Legal Documents; Mauro Dragoni, Serena Villata, Williams Rizzi, Guido Governatori; 2017.
11. Eunomos, a legal document and knowledge management system for the Web to provide relevant, reliable and up-to-date information on the law; Guido Boella, Luigi Di Caro, Llio Humphreys, Livio Robaldo, Piercarlo RossiLeendert van der Torre; 2016.
12. TULSI: an NLP system for extracting legal modificatory provisions; Leonardo Lesmo, Alessandro Mazzei, Monica Palmirani, Daniele P. Radicioni; 2012.
13. Miller G. A. WordNet: a lexical database for English //Communications of the ACM. – 1995. – T. 38. – № 11. – C. 39-41.
14. Manning C. et al. The Stanford CoreNLP natural language processing toolkit //Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations. – 2014. – C. 55-60.
15. McCune W. Release of prover9 //Mile High Conference on Quasigroups, Loops and Nonassociative Systems, Denver, Colorado. – 2005.
16. Martín Chozas P. Towards a Linked Open Data Cloud of language resources in the legal domain: дис. – ETSI_Informatica, 2018.
17. Kilgarriff A. et al. The Sketch Engine: ten years on //Lexicography. – 2014. – T. 1. – № 1. – C. 7-36.
18. Boella G., Di Caro L., Robaldo L. Semantic relation extraction from legislative text using generalized syntactic dependencies and support vector machines //International Workshop on Rules and Rule Markup Languages for the Semantic Web. – Springer, Berlin, Heidelberg, 2013. – C. 218-225.
19. Boella G. et al. Linking legal open data: breaking the accessibility and language barrier in european legislation and case law //Proceedings of the 15th International Conference on Artificial Intelligence and Law. – ACM, 2015. – C. 171-175.
20. G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. Commun. ACM, 18:613–620, November 1975.
21. L. Lesmo. The Turin University Parser at Evalita 2009. Proceedings of EVALITA, 9, 2009.
22. John A. K. et al. Legalbot: a deep learning-based conversational agent in the legal domain //International Conference on Applications of Natural Language to Information Systems. – Springer, Cham, 2017. – C. 267-273.
23. Hochreiter S., Schmidhuber J. Long short-term memory //Neural computation. – 1997. – T. 9. – № 8. – C. 1735-1780.
24. Van Opijnen M., Santos C. On the concept of relevance in legal information retrieval //Artificial Intelligence and Law. – 2017. – T. 25. – № 1. – C. 65-87.
25. Livermore M. A. et al. Law Search as Prediction //Virginia Public Law and Legal Theory Research Paper. – 2018. – № 2018-61.
26. Gain B. et al. IITP in COLIEE@ ICAIL 2019: Legal Information Retrieval using BM25 and BERT. – 2019.
27. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //arXiv preprint arXiv:1810.04805. – 2018.

28. Sun B. Information Structure Parsing for Chinese Legal Texts: A Discourse Analysis Perspective //International Journal of Technology and Human Interaction (IJTHI). – 2019. – Т. 15. – № 1. – С. 46-64.
29. Shaffer R., Mayhew S. Legal Linking: Citation Resolution and Suggestion in Constitutional Law //Proceedings of the Natural Legal Language Processing Workshop 2019. – 2019. – С. 39-44.
30. Osipov G. S., Smirnov I. V., Tikhomirov I. A. Relational-situational method for text search and analysis and its applications //Scientific and Technical Information Processing. – 2010. – Т. 37. – № 6. – С. 432-437.
31. Daniil Larionov, Artem Shelmanov, Elena Chistova and Ivan Smirnov. Semantic Role Labeling with Pretrained Language Models for Known and Unknown Predicates // Proceedings of Recent Advances of Natural Language Processing, 2019, pp. 620-630.
32. Sochenkov I. et al. Exactus like: Plagiarism detection in scientific texts //European conference on information retrieval. – Springer, Cham, 2016. – С. 837-840.
33. Hoekstra R. et al. The LKIF Core Ontology of Basic Legal Concepts //LOAIT. – 2007. – Т. 321. – С. 43-63.
34. Voronina I. E., Pigalkova E. A. Creating a basic ontology for the Russian legal system based on the LKIF-Core ontology // Bulletin of the Voronezh State University. Series: System Analysis and Information Technology. – 2010. – № 1. – p. 154-159.