Жизнен цикъл на извличане на именувани обекти в правната област

Named Entity Extraction Lifecycle in the Legal Domain

Станимира Йорданова[1]

Абстракт

Извличането на именувани обекти от правни текстове представлява предизвикателство в областта на правната лингвистика и изкуствения интелект. Правният текст се отличава със сложна структура, специфична терминология и голям обем, което налага прилагане на методи за обработка на естествен език за обработка с цел анализ и извличане на важна информация от неструктурираните данни. Целта на доклада е да проучи жизнения цикъл на извличане на именувани обекти от правни документи като изследва етапите на идентифициране и класифициране на важни за правната област обекти в текста, както и да анализира различните подходи на извличане на правна информация.

Ключови думи: Обработка на естествен език, извличане на именувани обекти, правни документи.

Abstract

Extracting named entities from legal texts is a challenging task in the field of legal linguistics and artificial intelligence. Legal text is characterized by a complex structure, specific terminology, and a large volume, which needs the application of natural language processing methods for processing, analyzing, and extracting important information from unstructured data. The goal of this paper is to analyze the life cycle of extracting named entities from legal documents by examining key information for extraction, the process and the challenges of identifying and classifying the objects to be extracted from the text.

Keywords: Natural language processing, Named Entity extraction, Legal documents

JEL: C880

Въведение

Извличането на именувани обекти (Named Entity Extraction) е специфична задача в извличането на важна информация от текстови документи, която цели идентифициране и извличане на конкретни реални обекти от документите, принадлежащи към предварително определени категории [1]. Именуваният обект е обект от реалния свят, който може да бъде обозначен със собствено име, например лица, местоположения, организации, дати и др.). Извличането на именувани обекти позволява свързване на обектите в текста с действителното им значение, което улеснява обработката на неструктурирания текст и подпомага по-доброто разбиране и тълкуване на текста в конкретната област на документа.

В правната област съществува голямо разнообразие от видове правни документи като договори, закони и подзаконови нормативни актове, съдебни актове, научни статии и др.). Правните документи често са дълги и сложни, съдържащи правна терминология, специализиран жаргон и нюансирани

[1] Гл. асистент д-р, Катедра „Информационни технологии и комуникации“, Факултет „Приложна статистика и информатика“, УНСС, e-mail: syordanova@unwe.bg

фрази, които могат да варират значително не само по юрисдикция, но и по контекст, което прави точното идентифициране на именуваните обекти предизвикателна задача [2]. Извличането на именувани обекти от правни документи помага за идентифицирането на участващи страни, дати, правни препратки и термини, които са от съществено значение за разбирането на последиците от правните текстове [3]. Извличането на важна информация от правния текст подобрява ефективността на анализа на правни документи, като се справя с предизвикателствата, породени от огромното количество неструктурирани данни в правната област.

Сложността на езика, използван в правните текстове, поставя значителни предизвикателства пред системите за обработка, анализ и извличане на полезна информация от правните документи и подчертава необходимостта от следване на структуриран подход при разработване на адаптивни модели за разпознаване на именувани обекти, който да гарантира качеството и ефективността на извличане на значима информация в правната област.

Целта на доклада е да се проучи жизнения цикъл на извличане на именувани обекти в правните документи, както и да се идентифицират значими обекти за извличане от правния текст, да се сравнят различни методи за анотиране и извличане на именувани обекти и да се очертаят основните предизивикателства в в този процес.

Извличане на именувани обекти от правни документи

Извличането на именувани обекти в правната област е процес, в който се прилага обработка на естествен език и позволява автоматизирано разбиране на правния език и трансформиране на неструктурирания текст в структурирани данни за аналитични цели. В този процес се идентифицират думи и фрази в текста, които представляват реални обекти и се класифицират в предварително определени категории, представляващи техните наименования. Проблемът, който се решава при извличането на именувани обекти е класификационен и налага обучение на класификационен модел, който да е способен да идентифицира в текста важните обекти и да ги категоризира в предварително определените категории.

Обучените модели с общо предназначение за извличане на именувани обекти от текст идентифицират обекти в следните категории – лице, организация, местоположение, дата, време, процент, стойност. В примера на фиг. 1 е представен резултат с извлечени именувани обекти на част от примерен текст в договор. Примерът е реализиран с обучен модел на английски език, наличен в Spacy и визуализиран с демо версията на displaCy Named Entity Visualizer [4].

Фигура 1: Пример за извличане на именувани обекти

Обучените модели с общо предназначение за извличане на именувани обекти обикновено се фокусират върху идентифициране на обектите към ограничен списък от категории. Специфичната структура на правните документи, която често е йерархична с множество препратки, и правната терминология изискват идентифициране на нови категории на обекти за извличане. Това налага разработване на модели за извличане на именувани обекти, специално адаптирани за конкретната правната област.

Жизнен цикъл на извличане на именувани обекти от правни текстове

Жизненият цикъл на извличане на именувани обекти обхваща поредица от взаимосвързани етапи на идентифициране и категоризиране на важните обекти в правните текстове с цел да се гарантира, че извлечената информацията е контекстуално правилна и да се подпомогнат правните експерти в анализа и взимането на правилни решения.

Процесът обхваща следните етапи Събиране на данни и предварителна обработка, Анотация на данните, Подготовка на данните, Обучение на модела, Оценка на разработения модел, Внедряване на модела и Непрекъснато подобрение на разработения модел (фиг. 2).

Фигура 2: Жизнен цикъл на извличане на именувани обекти

1. **Събиране и предварителна обработка на данните**

Етапът на Събиране и предварителна обработка на данни цели да гарантира разнообразието, качеството и количеството на данните. Необходимо е да се съберат различни правни документи от конкретна правна област, за да се създаде изчерпателен набор от данни, който отразява сложността на правния език. Качеството на данните се осигурява при предварителната обработка на данните. Събраните документи се подлагат на почистване и нормализиране, като премахване на шум, несъответствия, грешки и неподходяща информация и стандартизиране на формати (напр. малки букви, премахване на шумови думи). Прилагат се методи на обработка на текста като токенизация, маркиране на части от речта, автоматичен морфологичен (стеминг или лематизация) и синтактичен анализ с цел подобряване на точността на разпознаване на именувани обекти. Тези стъпки помагат за разбиването на текста на управляеми части, свеждането на думите до техните основни или коренни форми и идентифициране на тяхното граматично значение. *Токенизацията* разделя текста на отделни думи и елементи, което позволява на модела да разбира елементите на текста. *Маркирането на части от речта* предоставя синтактичен контекст към обектите като идентифицира граматичната категория на всяка дума в изречението, позволявайки на модела да прави разлика между видовете обекти. *Морфологичният анализ* (стеминг или лематизация) извършва намаляване на думите до техните основни или коренни форми, което допълнително подобрява способността на модела да идентифицира обекти в различни граматични форми. *Синтактичният анализ* помага за разбиране на синтактичната и граматичната структура на текста и по-добрата интерпретация на връзките между думи и фрази в изречението.

2. **Анотация на данни**

Етапът на анотация на данни включва добавяне на категория на важни обекти в текста като имена на страните, местоположение, адрес, дати, правни термини и др. Анотацията на данни осигурява данни с предварително известни категории на обектите, с които моделът да се обучи.

- *Подходи за анотация на именувани обекти в текста*

Използват се различни подходи за анотиране на данни, включително ръчна анотация, анотация, базирана на правила, полуавтоматизирани подходи, полуконтролирано и активно обучение, краудсорсинг и итеративна анотация [5]. Таблица 1 представя сравнение между подходите за анотиране на данни за извличане на именувани обекти.

Подход	**Описание**	**Предимства**	**Предизвикателства**
Ръчна анотация	Анотацията се извършва ръчно от правни експерти	Висока точност и разбиране контекста	Изисква много време и ресурси
Полуавтоматизирана анотация	Автоматизирани инструменти се комбинират с ръчно анотиране	Висока точност и ускорява процеса на анотиране Предпочитана за големи проекти	Зависимост от началните анотирани данни
Анотация с правила	Анотация, която използва предварително дефинирани правила за идентифициране на именувани обекти, чрез езикови характеристики, синтактични структури или специфични за областта знания	Ефективна в области с добре дефинирани типове обекти (позволява последователни и повтарящи се резултати в набори от данни) Правилата са лесни за разбиране и интерпретиране от моделите за машинно обучение, което улеснява идентифициране и коригиране на грешки	Не разпознава вариации в текста и не идентифицира нови или непознати обекти, които не са обхванати от съществуващите правила
Полуконтролирана анотация	Модел за анотация, който използва данни с категоризирани и некатегоризирани обекти	Подходяща при наличие на данни с некатегоризирани и категоризирани обекти Ефективна за адаптиране на модел, обучен в една област, към друга област	Зависимост от началните категоризирани данни и качеството им Изисква изчислителни ресурси, особено когато се работи с големи набори от данни Затруднена оценка на качеството на модела
Активно учене	Моделът за анотация активно избира най-информативните некатегоризирани обекти, които да бъдат категоризирани от експерт	Намалява ръчната анотация Подобрява ефективността на разработвания модел Използва голямо количество некатегоризирани данни Ефективна за адаптиране на модел, обучен в една област, към друга област	Изисква много време и изчислителни ресурси за големи набори от данни Качеството на модела зависи от избора на алгоритъм и качеството на данните
Краудсорсинг	Възлагане на задачата за анотиране на обекти на голям брой хора, често чрез онлайн платформи (напр. Amazon Mechanical Turk, MTurk)	Бързо и ефективно анотиране на големи набори от данни	Труден контрол на качеството на анотиране Анотирането може да изисква обучение и специфично за областта знание

Итеративна анотация	Многократно анотиране с цел подобряване на качеството на анотациите и модела за машинно обучение	Ефективна при сложни или с двусмисленост данни Помага за намаляване на аномалиите в категоризираните данни чрез идентифициране и коригиране на грешки Може да доведе до по-задълбочено разбиране на основните концепции и модели в данните	Изисква много време Може да изисква специфично за областта знание при идентифициране и коригиране на грешките Качеството на първоначалните анотации може да повлияе на цялостната точност на модела

Таблица 1: Подходи за анотиране на данни за извличане на именувани обекти

- *Схеми за анотиране на именувани обекти в текста*

Именуването на обекти в текста се осъществява чрез прилагане на схеми за анотация на текст, които използват от два до повече тагове, идентифициращи границите на именувания обект, когато се състои от повече от един токен (дума).

Често използвани схеми за анотиране на именувани обекти са IO (Inside-Outside), IOB (Inside-Outside-Beginning), IOE (Inside-Outside-End) и BIOES (Beginning-Inside-Outside-Ending-Single) [6].

Основната схема е IO**,** при която се присвояват към думите два тага - вътрешен таг (I) и външен таг (O). Тагът I е за именувани обекти, а тагът O е за неименувани думи. Например, [I-Person] е таг за именуван обект от категорията човек. Недостатък на тази анотация е, че не може да различи последователни имена на обекти от един и същи тип. Този недостатък може да се преодолее чрез използване на други две схеми IOB и IOE. Първата схема за анотиране въвежда трети таг към IO - начало на известен именуван обект (B). Останалите два тага - вътрешен таг (I) за именувания обект и външен таг (O) за неименувани думи - се запазват. Например, “New York” би получил следните тагове: New [B-City], York [I-City]. Втората схема, IOE, използва отново три тага - край на именуван обект (E), вътрешен таг (I) за именувания обект и външен таг (O) за неименувани думи. Например, “New York” би получил следните тагове: New [I-City], York [E-City]. BIOES схемата съчетава тагирането на IOB и IOE и добавя още един таг (S) за обект, който се състои от един токен. Останалите схеми за анотиране - BI, IE и BIES се фокусират върху тагиране на токени, които не са именувани обекти.

Изборът на схема за анотиране на обекти в текста зависи от данните, езика на текста и конкретното приложение за извличане на именувани обекти [7].

3. Подготовка на данните

Подготовка на данните за модела обхваща структуриране и избор на подходящи именувани обекти, които участват в обучението на модела с цел точно идентифициране на именувани обекти. Структурирането на текста обхваща векторно представяне на думите като разстоянието и посоката между векторите отразяват сходството и връзките между съответните думи, позволявайки на алгоритмите за машинно обучение да разбират и обработват семантичните и синтактични връзки между думите.

4. Обучение на модела

Този процес обхваща избор и прилагане на метод и алгоритъм за обучение, както и подход за организиране на процеса на обучение. Методите за обучение на модела за класификация на именувани обекти могат да бъдат категоризирани в следните групи (фиг.3): методи, използващи

правила; методи, използващи речници; методи, прилагащи машинно обучение; методи, прилагащи дълбоко обучение и хибридни методи [8].

Фигура 3: Методи за разработване на модела за класификация на именувани обекти

- **Методите, използващи правила**, класифицират именуваните обекти чрез предварително дефинирани правила за именуване на обектите. Тези методи се прилагат, когато обектите са ясно дефинирани чрез правилата и представляват предизвикателство, когато се прилагат към различни типове обекти или големи набори от данни, защото не могат да класифицират обекти, които не са обхванати от правилата.
- **Методите, използващи речник,** класифицират именуваните обекти чрез речник, включващ предварително именувани обекти. Тези методи ефективно се прилагат, когато речникът с обекти е необходимо да бъде статичен. Идентифицирането на нови обекти изисква редовна актуализация на речника с именувани обекти.
- **Методите, използващи машинно обучение,** класифицират именуваните обекти чрез обучение на данните с алгоритми за класификация на обектите в текста. Тези методи са универсални и могат да бъдат съобразени с различни области, при условие че има достатъчно количество данни за обучение.
- **Методите, използващи дълбоко обучение,** класифицират именуваните обекти чрез методи като двупосочен LSTM (BiLSTM) и модели, базирани на трансформарс (BERT и RoBERTa), които могат да улавят сложни връзки и контекстуални нюанси в текстовите данни. Тези методи могат да бъдат съобразени с конкретни области, но често изискват значително количество предварително анотирани данни за обучение.
- **Хибридните методи** комбинират методи с правила, речници и машинно обучение за класифициране на именуваните обекти. Тези методи могат да се прилагат в различни области. Тяхното внедряване и поддръжка изискват значително ниво на опит и знания по отношение на приложението им, значителни изчислителни ресурси и класифицирани данни за обучение.

5. Оценка на разработения модел

Оценката на качеството на модела за извличане на именувани обекти има за цел да сравни резултатите от класификацията на именуваните обекти, извършена от разработения модел, и първоначалната класификация на обектите в набора от данни за обучение. Тъй като извличането на именувани обекти е класификационна задача за решаване, в оценката на модела се използват показатели като Accuracy, Precision, Recall, F-score на ниво именуван обект.

6. Внедряване и непрекъснато подобрение на модела

Внедряването на модела обхваща интегриране на обучения модел в правни приложения (приложения за извличане на правна информация, обобщаване на правни документи и др.).

Непрекъснато подобрение включва събиране на обратна връзка за работата на модела и извършване на необходимите корекции и обучение на модела с допълнителни данни или подобрени техники, както и адаптиране на модела за обработка на различни видове правни документи.

Обучени модели за извличане на именувани обекти в правната област

Обучените модели представляват готови модели, които са обучени върху огромно количество данни и имат възможност да бъдат допълнително прецизирани и адаприрани за решаване на конкретни задачи. Предварително обучените езикови модели е необходимо да се настроят с цел да се приведат в съответствие с изискванията на новата задача, което води до подобрена производителност и ефективност. Първите обучени езикови модели са от 90-те години на миналия век. През 2017г. бе представен модел, базиран на невронна мрежа - Transformer, който значително подобри обработката на естествен език [9], а през 2018г. Google разработи друг обучен езиков модел, наречен BERT, който използва двупосочен енкодер за генериране на контекстуализирани представяния на думи в изречение [10]. И двата модела се превърнаха в стандартни компоненти на много системи за обработка на естествен език. През 2019г. и 2020г. излизат три варианта на BERT, които използват различен подход към обучението – RoBERTa, разработен от Meta (Facebook), XLNet, разработен от Google и GPT, разработен от OpenAI.

Тъй като предварително обучените езикови модели могат да бъдат адаптирани за решаване на конкретна задача, обучени модели за обработка и анализ на правен текст са [11,12, 13, 14]:

- LegalNLP - разработен за бразилски и португалски език и обучен със съдебни документи.
- Legal-BERT – разработен за английски език и обучен с европейски, английски и американски законодателни документи.
- Legal-RoBERTa и LexLM – разработен на английски език, обучен с LeXFiles корпус, включващ 11 отделни подкорпуса, които обхващат законодателството и съдебната практика от 6 англоезични правни системи.

Предизвикателства в извличането на именувани обекти от правни документи

Основните предизвикателства в извличането на именувани обекти от правни документи са свързани със спецификата на текста и езика, количеството и качеството на данните и методите на извличане.

Правните документи се отличават със специфична терминология и изказ, двусмисленост на думите, нюанси, жаргон и подразбиращи се значения, както и сложни структури на изреченията, което затруднява тяхната обработка. Тези предизвикателства могат да се преодолеят чрез разработване на модели за обработка на естествен език, съобразени с особеностите на правната област и които да помогнат за правилното тълкуване на правните термини, както и извличане на контекста, в който са използвани. Интегрирането на специфични за областта знания в извличането на именувани обекти може да доведе до по-точни модели на разпознаване на обектите, тъй като това би позволило на алгоритмите да разберат по-добре контекста около правните термини и техните последици в конкретни случаи [15].

Липса на стандартен формат в правни документи, чиято структура и формат могат да бъдат различни, затруднява моделите да обобщават и извличат ефективно обекти и концепции в различни типове документи. Графите на знанието (Knowledge Graph) могат да осигурят семантично представяне на правните понятия и техните взаимоотношения. Това помага за разбирането на контекста и премахването на неяснотите на обектите, които биха могли да бъдат объркващи поради различни формати. Графите свързват обекти в мрежа на знанието, като по този начин могат да бъдат по-лесно идентифицирани и категоризирани, ако се появяват в различни формати или контекст. Освен това, графите могат да разкрият скрити връзки и да дадат по-задълбочена представа за правните текстове, както и могат да бъдат обновявани с нови правни термини, което осигурява актуалност и точност на извличането.

Важен проблем в извличането на именувани обекти от правни текстове е наличието на ограничени данни за обучение на моделите за извличане, които изискват големи обеми от данни с високо качество и предварително класифицирани. Използването на големи езикови модели в анотирането на данни (GPT, BERT) като инструменти за създаване на анотирани данни, подобни на аннотираните от експерти, е възможност за справяне с проблемите с недостига на данни [16].

Заключение

Предизвикателствата в извличането на именувани обекти от правния текст са свързани както със спецификата на правния текст и правния документ, и избора на методи и инструменти за разработване на модели за извличане на именувани обекти, така и с приложимостта на резултатите от извличането.

Следването на етапите за извличане на именувани обекти осигурява по-добро качество на извличане на значима информация в правната област и приложимост на резултатите с цел да подпомогне правните експерти във взимането на важни решения. Итеративният характер на жизнения цикъл позволява непрекъснато усъвършенстване и адаптиране на разработените модели към променящите се изисквания и предизвикателства на различни правни области.

Преодоляването на предизвикателствата се подпомага от възможностите, които предоставят предварително обучени езикови модели, които могат допълнително да се настройват и адаптират за извличане на именувани обекти в конкретна правна област като осигуряват бързо и точно разпознаване на именувани обекти от правния текст. Тъй като тези технологии продължават да се развиват, се очаква те да играят ключова роля в трансформирането на традиционните правни работни процеси, позволявайки на правните експерти да се съсредоточат повече върху качество на правния анализ.

References

1. Maged, Refaat., Ahmed, Rafea., Nada, Gaballah. (2023). Named Entity Recognition From Biomedical Data. Doi: 10.1109/csci62032.2023.00141
2. Badji, I. (2018). Legal entity extraction with NER Systems. https://oa.upm.es/51740/1/TFM_INES_BADJI.pdf
3. A., Radhika., N, K, Bhasin., Y., Ravi, Raju., K.N.V., Satyanarayana., I., I., Raj. (2024). Optimization of Natural Language Processing Models for Multilingual Legal Document Analysis. 1-6. doi: 10.1109/incos59338.2024.10527598
4. displaCy Named Entity Visualizer, https://demos.explosion.ai/displacy-ent
5. Keraghel, I., Morbieu, S., & Nadif, M. (2024). A survey on recent advances in named entity recognition. *ArXiv, abs/2401.10825*.
6. Alshammari, Nasser, and Saad Alanazi. (2021) “The impact of using different annotation schemes on named entity recognition.” Egyptian Informatics Journal 22.3 (2021): 295–302.
7. Chen, Maojian, et al. (2021) “A Novel Named Entity Recognition Scheme for Steel E-Commerce Platforms Using a Lite BERT.” CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES 129.1 (2021): 47–63.

8. Basra Jehangir, Saravanan Radhakrishnan, Rahul Agarwal (2023). A survey on Named Entity Recognition — datasets, tools, and methodologies, Natural Language Processing Journal, Volume 3, 2023, 100017, ISSN 2949-7191, https://doi.org/10.1016/j.nlp.2023.100017.
9. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017) Attention Is All You Need. arXiv: 1706.03762.
10. Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina (October 11, 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv:1810.04805v2
11. Polo, Felipe Maia, et al. (2021) "LegalNLP-Natural Language Processing methods for the Brazilian Legal Language." Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. SBC, 2021, https://arxiv.org/abs/2110.15709
12. LEGAL-BERT. BERT-based machine learning model trained on hundreds of thousands of legal documents, https://opensource.legal/projects/Legal_BERT
13. Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos (2020). LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.
14. Ilias Chalkidis, Nicolas Garneau, Catalina E.C. Goanta, Daniel Martin Katz, and Anders Søgaard. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development (2022). In the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Toronto, Canada, https://arxiv.org/abs/2305.07507
15. Naik, V., & Kannan, R. (2023). Legal Entity Extraction: An Experimental Study of NER Approach for Legal Documents.*International Journal of Advanced Computer Science and Applications*. https://doi.org/10.14569/ijacsa.2023.0140389
16. Zin, M.M., Nguyen, H.T., Satoh, K., Nishino, F. (2024). Addressing Annotated Data Scarcity in Legal Information Extraction. In: Suzumura, T., Bono, M. (eds) New Frontiers in Artificial Intelligence. JSAI-isAI 2024. Lecture Notes in Computer Science(), vol 14741. Springer, Singapore. https://doi.org/10.1007/978-981-97-3076-6_6